

Langham
PO 20801
85720

NSA TECHNICAL JOURNAL

SPECIAL LINGUISTICS ISSUE II

Handle With Care: Russian Transliteration	A. J. SALEMME	1
Mokusatsu: One Word, Two Lessons	H. G. ROSENBLUH	17
Machine Translation—Where Is It Today?	C. D. McDONALD	23
Employment of Military Linguists	JACOB GURIN	31
The Voynich Manuscript—"The Most Mysterious Manuscript in the World"	J. H. TILTMAN	41
Linguistic Canonical Forms: Tools for Analytic Progress	C. H. BUTCHER	87

THE National Security Agency, established by presidential directive, is the national authority for all United States communications security activities. Its headquarters building is located between Washington, D. C. and Baltimore, Maryland. The Agency conducts its own recruiting and employment programs to hire promising individuals from all sections of the country.

NSA offers interesting and useful careers in many of the intellectual disciplines. People trained in languages and linguistics find the Agency a congenial place in which to work and in which to increase their knowledge and use their skills. Those trained in certain modern languages and area studies may use their skills as basic tools for research in a number of analytical fields. A high degree of flexibility and ingenuity is required because of the varied nature of the work. Systematic accumulation of information, data examination, and the preparation of special reports are important products of this research. The ability to read in a wide variety of subjects and to translate into good English is important. Fluency in speaking is not essential, but knowledge of idiomatic and colloquial speech is desirable. Those with an exceptional facility in languages are often assigned immediately to linguistic duties or are selected for training in an entirely new language. Scientific linguists have an opportunity to engage in a variety of practical applications of linguistics in many complex assignments.

If you wish to learn more about employment opportunities at NSA write to: Chief, College Relations Branch, National Security Agency, Fort George G. Meade, Maryland 20755, Attn.: M321.

THE NSA TECHNICAL JOURNAL

SPECIAL LINGUISTICS ISSUE II

This special issue of the NSA Technical Journal offers a sampling of the scholarly output of this Agency's linguists. As the National center for our country's communications security activities, the National Security Agency has to deal with a wide range of technical problems. Many of these require the services of professionals who possess abilities in foreign languages or in scientific linguistics. A major purpose of this publication is to draw the attention of interested readers outside the Agency to the opportunity of a career with NSA, utilizing and enlarging their skills in foreign languages and linguistics.

Examples of the tasks to be performed include the following: translation or summarization in English of technical foreign language material; transcription of foreign speech, or English summarization; preparation of studies, reports, etc. based on material in one or more foreign languages; preparation of linguistic aids, such as glossaries or more specialized products of language analysis. Other applications of linguistic talent may be needed in such areas as speech communications, computer applications to linguistic tasks and, not least, teaching of subjects in the fields of foreign language or linguistics in the Agency's School. Specific assignments may be highly specialized and may include uniquely rewarding research projects.

The articles which follow will serve to show to some extent the variety of interests which attract NSA's linguists. All find their duties challenging; many consider them quite captivating.

The Guest Editor for this special issue of the *NSA Technical Journal* is Dr. James R. Nielson, Chief of the Department of Intelligence and Language, National Cryptologic School.

Handle With Care: Russian Transliteration

BY A. J. SALEMME

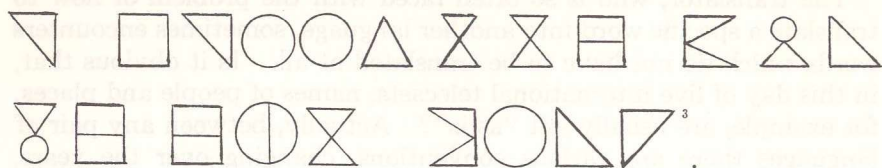
The translator, who is so often faced with the problem of how to translate a specific word into another language, sometimes encounters words which do not have to be translated at all. Is it obvious that, in this day of live international telecasts, names of people and places, for example, are usually left "as is"? Actually, between any pair of languages there are various conventions, changing over the years, which govern the treatment of names and other borrowed words. Accordingly there are various "traditional" treatments which sometimes defy simple logic. English speakers generally say "Rome, Naples" (not "Roma, Napoli"), but vacillate between "Leghorn" and "Livorno"; read "Louis XVI" as "Louis the Sixteenth" instead of the antique trade's "Louis Seize"; use "Rive Gauche" and "Left Bank" almost interchangeably; but seem to prefer "Champs Elysées, Moulin Rouge" to "Elysian Fields, Red Mill." Germans say "Mailand, Venedig" (not "Milan, Venice" or even "Milano, Venezia"), and the French say "Londres" (not "London"). But if the translator is aware of the relatively infrequent "traditional" treatments, he is usually safe in retaining "as is" personal and place names, as well as nouns and other words for which there is no precise equivalent in the language into which he is translating. "Guillaume, Colombey-les-deux-Eglises, sang froid" remain unchanged, instead of being translated, or almost-translated, as "William, Colombey-with-the-two-churches, cold blood." "Heinrich, Neustadt, gemütlich" are not "Henry, Newton, nice."

Of course, an additional problem arises when such words are borrowed from a language having a different writing system. Since, for example, Russian letters (like Japanese, Arabic, Burmese, or other non-Latin characters) are often completely unintelligible to the average American reader, or, even if they are intelligible, pose typographical difficulties, the words cannot remain entirely "as is"¹, but have to be transliterated into Latin letters. Thus, Иван Кузнецов,

¹ Except when some purpose is served by retaining the actual shape of the letter, as when an entire mystery novel hinges upon a single clue, such as a handkerchief which presumably belongs to the heroine ("The monogrammed H could only stand for 'Helen'") but proves on the last page to belong to the completely unsuspected Russian ex-countess Наталья Петровна.

Кривой Рог, солончак become "Ivan Kuznetsov, Krivoj Rog, solonchak" (but not "John Smith, Crooked Horn, salty soil")².

Unfortunately, transliteration is not as simple as sending in a quarter to the cereal company and being the first in the neighborhood to get a "Magic Message Decoder" with the beautifully one-for-one ability to read messages like:



What, then, are some of the problems making simple one-for-one transliteration from Russian difficult? In order to answer that

² Translation of personal and place names is by no means impossible, although the results range from the confusing to the ludicrous. Britishers, for example, who would be among the last to translate the name of a Frenchman "Pierre Boulanger" as "Peter Baker," apparently do not mind translating the first names of everyday Russians (calling them "Constantine, Peter, Nicholas" instead of "Konstantin, Petr, Nikolaj"). Perhaps this is an extension of the common British treatment of Russian historical names ("John Moneybags" instead of "Ivan Kalita," "John the Dread" instead of "Ivan IV," "Ivan Groznyj," or "Ivan the Terrible"). Nowadays, Russian last names are not translated by anyone, British or otherwise. True, Russian names often lend themselves to plays on words, such as that on "Scratchy" Tsarapkin's last name, but only a crude machine-translation program would actually translate Хрущев as "cock-chafers" or Борис Бойко as "Boris briskly/smartly" (the latter sounds somewhat like a Saki character, what?). Nor is mistreatment of place names by translating them impossible. There have been a few legendary instances such as the well-known flub of the German General Staff in misinterpreting a report from an agent in Spain to read that Churchill and Roosevelt would be meeting in "the White House," instead of where the man said, in "Casablanca." But most of the incorrect treatments of place names fall into the same internationally less critical class as the treatments by those machine-translation programs which, at their own risk, chop up into translatable chunks any new words the machine encounters. Item: the machine translation of Антикитера, the Russian spelling of Antikythera, a Greek island, as "antiwhaleparadise."

³ To the person just learning Russian, the Cyrillic letters may look as weird as the characters in the cereal box "codes." Accordingly he may hope to memorize the alphabet, with or without mnemonic devices like "Л has a hook like an L-ephant's trunk." His "Magic Message Decoder" easily changes Берлин and Лондон to "Berlin" and "London," confirming his guess that it's just one step from Russian spelling to English meaning. When Париж comes out "Parizh," he starts to wonder, and he soon finds out, with Черное Море, that Russian-spelling-to-English-meaning usually has to go through the step "Russian meaning:" "Chernoe, can that be right? More, what's that?"

question it is necessary first to invoke the names of Tschaikowsky, Chekhov, and Khrushchev⁴.

A person who does not know Russian sees no illogicality in the spellings of the familiar names Tschaikowsky and Chekhov. He simply assumes that the conventional spellings in Latin letters reflect some strange Russian spellings. He is so used to these conventional spellings that he rejects out of hand the suggestion that, since both Russian names begin with the Russian letter Ч one could easily throw away the *T*s in Tschaikowsky with no loss⁵. No, he feels, Chaikowsky and Chajkovskij "look funny"⁶ but Tschaikowsky "looks right." So does Chekhov, and so does Rachmaninoff. Even though the *w* in Tschaikowsky, the *v* in Chekhov, and the *ff* in Rachmaninoff all represent the same letter в. And even though the *kh* in Chekhov and the *ch* in Rachmaninoff represent the same Russian letter х⁷.

Are all those variations completely illogical? Actually, the person who knows Russian realizes that the differences in the Latinized spellings of Russian names on record albums and on book covers are usually the result of the fact that some Russian names might become better known in an "English" spelling (Kutuzov, Lobachevsky, Shostakovich), others in a "French" spelling (Koussevitzky, Mousorgsky, Chaliapin, Scriabine), and still others in a "German" spelling (Rachmaninoff, Tschaikowsky, Gabrilowitsch). These differences also explain the situation—often baffling to a person who does not know Russian—in which a person with the last name Vasilev has a

⁴ These names are apparently as ritualistically vital at the start of a discussion of Russian transliteration problems as the name of the Almighty is, when used to tag the last sentence of a political speech at an American nominating convention.

⁵ One of my music-loving friends feels that, while we're at it, we might just as easily throw away all of Tschaikowsky's music with no loss either, but that is a matter beyond the scope of this article.

⁶ So does the spelling of the identically pronounced Polish name Czajkowski. But since Polish is written in Latin letters, there is no question of choosing among various transliterations.

⁷ Incidentally, the Russian letter X, used by itself, is not equivalent to Latin X in the meaning "unknown." Instead, the Russian letter H is used in that sense (e.g., полковник H. = "Colonel X.") and the Russian letter X connotes a not too acceptable Russian noun, and its various affectionate, diminutive, augmentative, pejorative, and other derivative forms. Does that explain why Russian-speaking Americans, seeing references in the press to President Kennedy and Nikita Khrushchev as "the two K's," used to suggest calling the latter "Comrade X," and then grin at one another? Or why sweet young things who are asked to translate Russian texts likely to contain crude expressions are advised, "If you can't find it in the dictionary and it begins with x, don't ask!"

brother with what looks like the completely different last name Wassiljew.

The examples in the preceding paragraphs show how the process of representing Russian in the various languages written in Latin letters is complicated because the individual languages have their own spelling conventions (Russian *y* is usually transliterated into English as *u*, but into French as *ou*). In addition, inconsistencies develop between transliteration (the representation of a word in a different language by means of an alphabetic, letter-by-letter conversion) and transcription (the representation of a word in a different language by means of a more or less phonetic, sound-by-sound conversion). Cf. Потемкин = Potemkin or Potyomkin. The process of transcription, in its turn, is often complicated by the lack of completely equivalent sounds in the various languages (e.g., the Russian letters *x* and *ы* which represent sounds that do not occur in English) and by subjective interpretations and preferences.

Inconsistencies in the representation of Russian words have a long history. The name "Moscow," for example, is a partial transliteration of Москва, but the seventeenth-century name "Muscovy" is obviously a transcription, as attested to by the accurate representation, in the first syllable, of the feature of *аканье* (pronunciation of unaccented *o*'s as *a*'s) typical, then as now, of the Moscow area.

Once established, these inconsistencies are hard to change, much less to "correct." By now, English-speaking people are used to the spelling Moscow, rather than Moskva, Maskva, Muskva, or Muscow. They are also used to seeing the name Tschaiakowsky in the German transliteration, but would probably not recognize it as listed in the Library of Congress files (Chaikovskii), just as they become used to the spelling Khrushchev but would probably not have been able to recognize the German spelling Chruschtschew.

Conversely, most Russians, long accustomed, from reading translations of James Fenimore Cooper's works, to the letter-by-letter transliteration томагавк ("tomahawk"), would not recognize a more accurate, but still less than perfect, transcription⁸.

The average translator of Russian is familiar enough with some other European language to understand why the name Пушкин, for example, assumes the following forms in the various European languages:

⁸ For one thing, what does Russian do with the sound represented by English *h*, a sound which is not present in standard Russian?

Pushkin	English	Poesjkin	Dutch
Pouchkine	French	Puskin	Hungarian
Pusckin	Italian	Puszkín	Polish
Puschkin	German	Puškin	Czech

Accordingly, he can usually explain, in terms of transliteration or transcription from Russian into English, French, or German, the seeming inconsistencies in the spelling of Russian words in English dictionaries such as the Webster's Third New International Dictionary⁹. Some of these words are cited in Webster's in one spelling only. Usually this is a spelling which corresponds, more or less, to ordinary English spelling conventions¹⁰, e.g.,

apparat	аппарат	Krivoi Rog	Кривой Рог
apparatchik	аппаратчик	Makeevka	Макеевка
artel	артель	Nikolaev	Николаев
babushka	бабушка	oblast	область
Balaclava	Балаклава	samogonka	самогонка
balalaika	балалайка	samovar	самовар
chastushka	частушка	Samoyed	Самоед
Chechen	Чечен	Saratov	Саратов
Cheka	Чека	solonchak	солончак
Chernovtsy	Черновцы	sputnik	спутник
chernozem	чернозем	taiga	тайга
Ivan	Иван	troika	тройка
kasha	каша	tundra	тундра
Kharkov	Харьков	Ulan-Ude	Улан-Удэ
Khlyst	Хлыст	vodka	водка
Kiev	Киев	Yaroslavl	Ярославль
kolkhoznik	колхозник	Yenisei	Енисей
Komi	Коми	Zhdanov	Жданов
Kremlin	Кремль		

⁹ Prerevolutionary and Soviet proper names and other Russian words have, for the most part, been borrowed directly from the Russian or through French and German. An occasional Russian word may have been borrowed through a European language other than the three mentioned (e.g., "czar," borrowed through the Polish spelling). But so far as I know, there is no large body of Russian words in Webster's spelled, say, in Dutch or Hungarian. I do, however, have a phonograph record cover, written in English, which lists Russian song titles such as "Nje rewnoej" and "Sjewodnja La Loebloe." These titles remain unintelligible until one reads the comment that the record was made by Russian emigres now living in the Netherlands. Then it is easy to recognize the (sometimes garbled) Dutch spellings of Russian words.

¹⁰ In these conventional English spellings, *и* and *й*, and sometimes *ы*, are rendered as *i*, *e* is rendered as *e* or *ye*, *к* is sometimes rendered as *c*, *х* and *ж* are rendered as *kh* and *zh*, and *ь* is ignored.

Occasionally the single spelling in Webster's may be in accordance with French transliteration or may include an attempt to represent a palatalized Russian consonant, e.g.,

knout ¹¹	КНУТ
Tchebycheff inequality	Чебышев
soviet	совет

But usually, in a spirit of "you-pays-your-money," Webster's lists two or more spellings per Russian word. Some of these sets of variants are understandable as alternate English/French/German spellings, but others are understandable only as various ways, traditional or otherwise, of getting the word into something like intelligible English. Sometimes Webster's throws all the variants of a Russian word into one steaming pot from which, on the end of a long stick, the user is supposed to fish out his preference. Although Webster's gives the definition under its preferred form (the first word in each set of variants below), that preference is not always consistent. E.g.,

Archangel, Arkhangelsk	Архангельск
Belorussian, Bielorussian, Byelorussian	Белорусский
beluga, belouga	белуга
borsch, borscht, borsht, bortsch, borshch	борщ
caftan, kaftan	кафтан
Cheremis, Cheremiss	Черемис
chervonets, chervonetz, tchervonets, tchervonetz	червонец
chetvert, tchetvert	четверть
chevkinite, tscheffkinite	чевкинит
guberniya, gubernia	губерния
intelligentsia, intelligentzia,	интеллигенция
jarovize, iarovize, yarovize	яровизировать
Kirghiz, Kirghese, Kirghis, Kirgiz, Khirghiz ¹²	Киргиз
kok-saghyz, kok-sagyz	кок-сагыз
kolkhoz, kolkos, kolkhos	колхоз

¹¹ Pronounced /'naut, 'nüt/. The entry "knut" /'(kə)'nɒt/ is something else entirely ("a fop of the late 19th and early 20th century").

¹² So where's "Khirghizh"?

Korsakoff's psychosis, Korsakow's psychosis¹³
koumiss, kumiss, kumys, kumyss
krym-saghyz, krim-saghyz
muzhik, muzjik, mujik, moujik
piroshki, pirojki
politburo, politbureau
polynya, polynia
sastruga, zastruga
Sebastopol goose
Sevastopol
shashlik, shashlick, shaslick
solonetz, solonets
sovkhoz, sovkhos, sovhoz
technicum, technikum
tovarich, tovarish
ukase, ukaz
uvarovite, ouvarovite, uwarowite
Uzbek, Uzbeg, Uzbak, Usbek, Uzbeg
Yerevan, Erevan, Erivan
zemmi, zemni

Корсаков
кумыс
крым-сагыз
мужик
пирожки
политбюро
попыньа
заструга
Севастополь
Севастополь
шашлык
солонец
совхоз
техникум
товарищ
указ
уваровит
Узбек
Ереван
земный

The inconsistencies in the English spelling of Russian words in Webster's could actually be worse, since there are at least 15 Russian transliteration systems currently in use at various governmental and private institutions, libraries, research projects, etc., in the United States, Great Britain, France, and Germany alone. The following table shows a few of the most common ones. It could easily have been made even more complex by including transliteration systems used at individual libraries such as the New York Public Library, and individual erratic transliterations based on the pronunciation, rather than the spelling (e.g., "e after consonant = ye, as in нет = nyet; or кс = x, as in Алексей = Alexei). The dashes in the table indicate that the transliteration in the particular system is the same as in the BGN system.

¹³ "After Sergei Korsakov."

BGN ^a	LC ^b	NSA ^c	AATSEEL ^d	MT ^e	French ^f	German ^g
а	a	—	—	—	—	—
б	b	—	—	—	—	—
в	v	—	—	—	v, ff ^h	w, ff ⁱ
г	g	—	—	—	g, gu ^j , gh ^j	—
д	d	—	—	—	—	—
е	e, ye ^k	e ^l	e ^l	e	e, ie	e, je
ё	ë, yë ^k	e ^l	e ^l	ë	yo	jo
ж	zh	—	—	ž	j, g	sh
з	z	—	—	—	—	z, s
и	i	—	—	—	i, i ^m	—
й	y	i ⁿ	j	j	i	j
к	k	—	—	—	k, c	—
л	l	—	—	—	—	—
м	m	—	—	—	—	—
н	n	—	—	—	n, ne ^o	—
о	o	—	—	—	—	—
п	p	—	—	—	—	—
р	r	—	—	—	—	—
с	s	—	—	—	s, ss ^p	s, ss ^q
т	t	—	—	—	—	—
у	u	—	—	—	ou	—
ф	f	—	—	—	—	—
х	kh	—	—	x	—	ch
ц	ts	ts ⁿ	—	c	—	z
ч	ch	—	—	c	tch	tsch
ш	sh	—	—	w	ch	sch
щ	shch	—	—	5	chtch	schtsch
ъ	"	—	—	7	y	y
ы	y	—	—	—	—	—
ь	'	—	—	6	y ^a	j
э	e	ë	eh	3	eh	e
ю	yu	iu ⁿ	—	h	you, iou	ju
я	ya	ia ⁿ	—	4	ya, ia	ja

^aU. S. Board on Geographic Names. Most U. S. government agencies adhere to the ruling given in the *GPO Style Manual* that "The spelling of geographic names must conform to the decisions of the U. S. Board on Geographic Names." For purposes of consistency, those agencies also use the BGN system for transliterating Russian words other than geographic names.

^bLibrary of Congress. This is one government agency which does not use the BGN transliteration system.

^cNational Security Agency. Like the Library of Congress, the National Security Agency is authorized to use its own Russian-to-English transliteration system. NSA does, however, use the LC system for library accessioning.

^dAmerican Association of Teachers of Slavic and East European Languages. System based on Czech alphabet.

^eSystem devised for machine-translation (MT) research by Georgetown University and subsequently adopted by IBM for the same purpose. Each Russian alphabetic symbol is unambiguously represented by a single Latin letter or arabic numeral. (Ambiguity between use of arabic numeral to represent a Russian alphabetic symbol and use as an actual numeral is reconciled by assigning a code prefix to actual numerals.)

^fCombination of common Russian-to-French transliteration systems.

^gCombination of common Russian-to-German transliteration systems.

^hWhen final (e.g., Уруссов, Ouroussoff).

ⁱWhen final (e.g., Киселев, Kisselyoff).

^jWhen followed by е or и (e.g., Тургенев, Tourgueneff, Дягилев Diaghileff).

^kWhen preceded by consonant. Otherwise (initial, or when preceded by vowel, ъ or ѣ) ye, ye. е and э indistinguishable in transliteration when preceded by consonant.

^lAlways e, regardless of whether initial or preceded by vowel, consonant, ъ or ѣ; no distinction in transliteration between е and ё.

^mWhen preceded by о (e.g., стоит, stoit); ий = y.

ⁿThe "modified" LC system omits diacritics; i, ts, iu, ia respectively.

^oWhen final and preceded by vowel (e.g., Ленин, Lenine).

^pBetween vowels (e.g., Киселев, Kisselyoff).

^qBetween vowels (e.g., Герасимов, Gerassimoff) or initial (Сомов Somoff).

^rый = y.

^sTransliterated when followed by vowel (e.g., Савельев, Savelyeff) ignored when followed by consonant or at end of word (e.g., Кольцов Гоголь, Koltsoff, Gogol).

With so many transliteration systems in existence, it obviously is not a question of which one is "right." It's analogous to deciding which is the "right" color to repaint the living room walls. First of all, what's wrong with the present color? Then, if they really *have* to be painted, wouldn't one of the premixed colors at the paint store be

close enough to what is desired? If not, how much does it cost to have colors custom-blended? Would changing the wall color require buying new carpets and new furniture for the living room? And the dining room? Would buying a new dining room set necessitate enlarging the dining room? Wouldn't knocking out the dining room wall mean chopping down the flowering crab just when it was beginning to look so pretty? Wouldn't buying a new fully-grown replacement from the nursery be better than buying one of those 2-foot sticks they sell at Safeway? Anyone who knows how these discussions usually end ("Let's just leave the walls seasick green!") can imagine how the Library of Congress, with its transliteration system nicely worked out and its Russian accessions neatly catalogued for a generation, felt when, toward the turn of the century, it became the law of the land that spellings of geographic names in U. S. government publications would be in accordance with decisions of the U. S. Board on Geographic Names. How many file cards to change? How many man-hours to train new people to change them? How many months to change how many cards? How many walls to knock out? ... Is it really surprising that, 60 years after the decision to standardize the spelling of geographic names in U. S. government publications, the Library of Congress lists Soviet place names at two separate locations in the catalogue when the LC transliteration differs from the BGN?¹⁴

That arrangement led to an amicable settlement between BGN and LC, so that nowadays one seldom hears remarks like "You and your crummy diacritical marks!" and "Oh yeah! Well, what about your crummy 'y' that you throw in when you can't think of anything better to do!" It also united LC and BGN spiritually when, still another generation later, another Johnny-come-lately government agency said, "A pox on both your transliteration systems!" And here we stand, in the Computer Age, with Russian words spelled every which way so that the poor dumb machine doesn't know which way to turn. And the walls are *still* painted that same seasick green.

One of the few virtues of strict adherence to any transliteration system is that it obviates the translator's necessity to cope with unrealistic rules like "If a Russian place name already has a standardized English spelling, that spelling should be substituted for the

¹⁴ The difference between Pervomaïsk (LC) and Pervomaysk (BGN) or between Ul'ianovskoe (LC) and Ul'yanovskoye (BGN) is perhaps slight enough not to confuse or inconvenience the user of the catalogue, but there is a very large number of catalogue cards and a considerable number of catalogue-user/feet between the spellings Erevan (LC) and Yerevan (BGN), or between Íáilo (LC) and Yaylo (BGN).

transliteration." Instead he can systematically ignore all such unsystematic "standardizations"¹⁵. Thus,

	BGN	NSA	But not
Архангельск	Arkhangel'sk	Arkhangel'sk	Archangel
Астрахань	Astrakhan'	Astrakhan'	Astrakhan
Горький	Gor'kiy	Gor'kij	Gorki
(Cf. Горки, Gorki)			
Куйбышев	Kuybyshev	Kujbyshev	Kuibyshev
Севастополь	Sevastopol'	Sevastopol'	Sebastopol

Unlike transliteration, which is the attempt to change the rendition of a word from one writing system to another writing system, transcription is the attempt to render in a particular writing system not the way a word looks, but the way it sounds¹⁶. Thus, transcription need not even involve a second language. It may be unscientific (e.g., thUs, tranz-KRIP-shun, LANG-uage), semiscientific (thūs, trānskrīp'-shūn, lāng'-gwīj), or scientific (ðʌs, trənskrɪpʃn, læŋɡwɪdʒ).

The problem of transcription (or "phonetic transcription") is made more complicated when a word from one language is being transcribed into another, since it is a rare phoneme that is exactly the same in any two languages. Obviously the solution that is closest to the ideal is a scientific transcription system that renders unambiguously all the meaningful sounds of all languages, not excluding Hottentot clicks or the Russian ы. This system, like the All-Purpose Universal International Transliteration Scheme, will probably never be achieved in this world. In the meantime the IPA (International Phonetic Association) system is as good as any. According to this system a French or Russian linguist will recognize in /ðʌs/, /trənskrɪpʃn/ and /læŋɡwɪdʒ/ the English words "thus," "transcription," and "language," and an American linguist will recognize in /al:ɔr/, /trānskripsjō/, and /lāga:ʒ/ the French words "alors," "transcription," and "langage," and will recognize in /tak/, /trənskrɪpʃjə/, and /jəzɪk/ the Russian words так, транскрипция and язык.

But words can also be transcribed from one language to another according to a less than scientific system. They often are, in fact, not only because typewriters lack the necessary diacritical marks and because print shops lack the necessary type fonts, but mostly because nonlinguists are scared away from strange-looking symbols. The English words "thus," "transcription," and "language" spelled in a

¹⁵ Except that everyone agrees that, despite the rules, it would be ridiculous to treat Москва and Киев as anything other than "Moscow" and "Kiev."

¹⁶ This is comparable to English shorthand or stenotypy, when, for example, "cats" and "Katz" are written with the same symbols.

system that is "easy-to-pronounce" for a Frenchman might look like "zasse," "tranne-SCRIPPE-chonne," "LINGUE-gouadzhe," and spelled to be "easy-to-pronounce" for a Russian might look like зас, трэнскрипшн, лэнгуэдж. Certainly there is little need to remind the reader that he has seen similar systems for making French and Russian "easy-to-pronounce" for the English-speaking student: "ah-LOHR," "trahn-screep-see-OHN," lawn-GAHZHE"; "tahk," "trahn-SKREEP-tsee-yah," "yah-ZOOIK."

The major problem that transcription poses in technical translation is recognizing when a non-Russian place name or personal name has been transcribed (in one system or another) instead of being transliterated (in one system or another). There being no logical pattern for treatment of foreign words borrowed from one language into another, non-Russian place names or personal names sometimes have two or more accepted variants in Russian. One variant might be a letter-by-letter transliteration that takes no consideration of the actual phonemes in the original language. By this method, Newton, Shakespeare, Shaw, and Harrow would be rendered Невтон, Шакэ-спэарэ, Схав, Гарров¹⁷. The other variant might be an attempt at a phonetic equivalent, by which method the same English names would be rendered Ньютон, Шекспир, Шоу, Хэрроу. Or the German city Hildesheim might be rendered as Гильдесгейм or Хильдесхайм.

Letter-by-letter transliterations such as Невтон are typical of an age when international communications were conducted mostly by means of the written word and French was the language of diplomacy. Transcriptions such as Ньютон are typical of the present age of electronic communications allowing the transmission of the actual pronunciation of the name or other word to be assimilated into another language. True, these pronunciations are often ignored by large segments of the population. The power of the printed word is still so great in determining the generally accepted pronunciation of foreign names and other words that one almost wishes that спутник were assimilated into English in the spelling "spootnik" (pronunciation used by one news announcer for every hundred that use a pronunciation rhyming with "nut pick"), or that Хрущев had been rendered in the American press as "Khrushchhoff" (which would at least have lessened the frequency with which one heard "Crew Chef").

At any rate, when foreign place names and personal names are cited in the daily press or in internationally oriented scientific literature, they often assume a form midway between transliteration and

¹⁷ You and I know that these spellings look ridiculous, but apparently Russians don't. A Russian-language newspaper published in New York, for example, used to run ads for the Аппов ("Arrow") farm.

transcription. For that reason non-Soviet place names and personal names appearing in Russian text can only rarely be transliterated by the same letter-by-letter procedure that is used for Soviet names. Резерфорд is not "Rezerford," but "Rutherford." Гомель is "Gomel" but Марсель is "Marseilles." Бойцов is "Bojtsov" but Кройценгеймер is "Kreuzenheimer."

If the translator of technical Russian encounters in Russian text place names in a language with which he is unfamiliar, he can usually find them on a map printed in Latin letters. By comparing that map with one printed in Cyrillic characters, he can discover, almost involuntarily, that Russian У, for example, is usually equivalent to French "ou" or Dutch "oe" (e.g., Тулуз, Гус, Toulouse, Goes). Maps can also be of great use in determining how to render non-Soviet personal names encountered in Russian text. By comparing Soviet and American maps of France, for example, one can easily construct a Russian-to-French conversion chart that can be of partial assistance in deciphering Russian transcriptions of French personal names¹⁸. A

ai Бодлер	Baudelaire	ée Добре	Daubrée
ai Леметр	Lemaître	ei Корнель	Corneille
ais Бомарше	Beaumarchais	er Оже	Auger
aix Сампе	Sampaix	es Дешан	Deschamps
ay Корде	Corday	et Бизе	Bizet
e Лепаж	Lepage	eu Пеккер	Pecqueur
é Абеляр	Abélard	eui Вернелль	Verneuil
è Ампер	Ampère	ez Дюмурье	Dumouriez
ê Ангулем	Angoulême		

conversion chart that would be more reliable than one reflecting the vagaries of French spelling might be one prepared from a comparison of Soviet and American maps of China, containing such equivalent values as ЧЖ = ch (Wade-Giles transcription), ч = ch', ин = ing, инь = in. The values contained in such a chart could then be used both for place names (ЧЖЭНДИН = Chengting, not Chzhendin), and for personal names Чжоу Энь-лай = Chou En-lai, not Chzhou Ehn'-laj).

By way of an exception, in order to facilitate library searching, non-Soviet personal names should be both transliterated and transcribed when they occur in Russian-language bibliographies. E.g.,

¹⁸ I'm afraid that this procedure is not as simple as it sounds. To get an idea of how complicated it is to convert Russian spellings of French names back into the original spellings, one need consider only the Russian vowel e and some (I have been fooled too many times to say "all") of the spellings it represents in French names.

Chzhou Ehn'-laj [Chou En-lai], "Speech at the Third Session of . . .," *Narodnyj Kitaj* [People's China], . . .

Uitmen, Uolt [Whitman, Walt], *List'ya travy* [Leaves of Grass], Moscow, 1955.

Names like Чжоу Энь-лай are instantly recognizable as being non-Soviet. For one thing, where's the -ович? But what if the name is that of a less well-known personality? Even then the translator will often have a clue, because usually the presence of two initials before a last name in Russian text indicates that the person referred to is or was a Russian citizen, e.g., Н. С. Хрущев, А. П. Чехов. But usually the presence of a single initial before a last name indicates a non-Russian, e.g., Ш. де Голль. Д. Эйзенхауэр.

While this is not an invariable rule, it is observed often enough to serve the translator as a general guide for the treatment of personal names. Russian names (pre- and postrevolutionary) are transliterated in accordance with the preferred transliteration table. But non-Soviet names should be returned to their original spellings in Latin letters or, for non-European languages, to a more or less standardized spelling in Latin letters.

Hence the difference in treatment between Soviet names like:

Г. П. Голыц	G. P. Gol'ts
А. М. Пешковский	A. M. Peshkovskij
М. Е. Тер-Минасян	M. E. Ter-Minasyan

and names like:

	Not	But	
А. Голыц	A. Gol'ts	A. Holz	(German)
Б. Шишковский	B. Shishkovskij	B. Szyszkowski	(Polish)
Ч. Кай-ши	Ch. Kaj-shi	Chiang Kai-shek	(Chinese)
С. Иосида	S. Iosida	S. Yoshida	(Japanese)

Names of well-known world personalities usually offer no difficulty to the translator, who automatically converts the spelling де Голль for example, to "De Gaulle," and, recognizing in the letter Ш. the French pronunciation of the digraph "ch," transcribes it as "C." instead of transliterating it as "Sh."

But when less well-known persons are mentioned in Russian text, the translator often has to do a certain amount of searching in bibliographical or other sources to find the actual spelling. If this search is unsuccessful, it is better for him to transliterate the name than to guess. The treatment "Dzh. Kelli [as in text]" is better than the bland assumption "J. Kelly" (since "G. Kelley" is as valid a guess). Another acceptable treatment for a name which the translator has

been unable to locate in its original spelling is "Dzh. Kelli [retransliterated from Russian]"¹⁹.

Incidentally, in at least two instances the initials of Soviet citizens require special treatment in translation. The first is when an author supplies his initials to indicate that he has added his own emphasis to a cited quotation. An article by K. I. Bakhman, for example, contains the sentence, "А. А. Реформатский утверждает, что грамматика любого языка может выражаться только этими способами (разрядка моя - К.Б.)" The author's initials after such expressions as разрядка моя ("my spacing") and подчеркнуто мной ("my italics," "italics supplied") should be suppressed in the English translation²⁰.

The second instance is when an author refers to himself in a context that balances delicately between the formal and the familiar. For example, the name of the author of an article written by a kindergarten teacher was given in a Soviet journal, in the ordinary formal manner, as А. К. Matveeva. In the article the author quoted several bits of conversations with the children, in which references were made to the cook and other persons by the familiar form of address with the first name and patronymic. But, describing another situation, the author quoted one child who began a question politely, Извините, А. К., но . . ." Obviously this creates a problem for the translator. He cannot guess what first name and patronymic the child actually used (Anna Konstantinovna? Aleksandra Kuz'minichna?). Neither can he translate it as, "Excuse me, А. К., but . . ." since that might create the impression that Russian kindergarten children speak like Madison Avenuers. The best he can do is translate the polite beginning as "Excuse me, Miss Matveeva, but . . ." (ignoring the possibility that she might be Mrs. Matveeva).

Personal names also occur occasionally as part of the names of Soviet institutions, which are named after people, famous or otherwise, by adding the word имени ("name of") and the person's name in the genitive case to the basic name of the institution (e.g., Московский государственный университет имени М. В. Ломоносова). The

¹⁹ This treatment is analogous to the treatment of translations into Russian of excerpts from books published originally in a different language, say French. A translation from Russian into English might look different from a translation from the original French into English. Similarly, a back translation of a Russian excerpt from an English-language source might look very different from the original English. If the translator can locate the original source of such excerpts, he can act appropriately. But if he cannot locate the source, it is desirable to indicate "[retranslated from Russian]".

²⁰ Note to the young translator: Enjoy this sensation. It isn't often you're allowed to, much less advised to, throw anything away.

translation treatment of such institutional names varies according to editorial preference. Some editors, attempting to achieve a form similar to that of English ("Walter Reed Army Hospital"), prefer treatments such as "M. V. Lomonosov Moscow State University." Others prefer keeping the same word order as in Russian, and either transliterating the word имени or rendering it by the arbitrary abbreviation "i/n" (for "in name of"): "Moscow State University imeni (or: i/n) M. V. Lomonosov."

Some editors object that moving the person's name to the front when translating long institutional names (treating, for example, Московский художественный академический театр СССР имени М. Горького as "M. Gor'kij Moscow Art Academy Theater of the USSR") produces names reminiscent of the "New Ashmolean Marching Society and Students Conservatory Band;" But what's so wrong with that? One achieves, in fact, a certain military snap by translating Краснознаменная, ордена Ленина и ордена Суворова 1-й степени военная академия имени М. В. Фрунзе as "M. V. Frunze Red Banner Order of Lenin and First-Degree Order of Suvorov Military Academy." (Although I'd hate to write the lyrics, in any language, for that academy's school song.)

Whether or not the translator moves the person's name to the front (check with your friendly editor), he should always remember to change any inflected form from the genitive case to the nominative. In addition, if a Soviet institution is named after a non-Russian, he should remember to spell the name appropriately, instead of transliterating it: Московский педагогический институт имени Карла Либкнехта, "Moscow Pedagogical Institute imeni Karl Liebknecht."

In addition to Soviet and non-Soviet personal and place names, certain other classes of words have to be transliterated instead of translated. These include various types of abbreviations, the names of certain Soviet chemical compounds, medicines, synthetic fibers, etc. The translation problems that these words pose could easily fill another article. Until that article hits the stands, here are a couple of teasers. When is СССР *not* the well-known abbreviation such as seen, among other places, on the sweatshirts of Russian lady shot-putters? If хлор is "chlorine," what's хлорин? Choose one: The preferred rendition of миарсенол (a Soviet antisypilitic compound) as "miarsenol" instead of "myarsenol" (a) *is* (b) *is not* coincidental to the fact that the word was not formed by mimicking a Cockney describing the ravages of the disease?

Mokusatsu: One Word, Two Lessons

BY HARRY G. ROSENBLUH

mokusatsu 黙殺^{xx}—suru, *v.* take no notice of; treat (anything) with silent contempt; ignore [by keeping silence]; remain in a wise and masterly inactivity.

—Kenkyusha's New Japanese - English Dictionary, p. 1129.

The story of how an ill-chosen translation of the Japanese word *mokusatsu* led to the United States decision to drop the world's first atomic bomb on Hiroshima is well known to many linguists. But perhaps it would not be amiss to retell it briefly just in case some reader of this essay is unfamiliar with the word—and in the hope that readers may be inspired to avoid the two tragic linguistic errors that the story points up.

In July of 1945 allied leaders meeting in Potsdam submitted a stiffly-worded declaration of surrender terms and waited anxiously for the Japanese reply. The terms had included a statement to the effect that any negative answer would invite "prompt and utter destruction." Truman, Churchill, Stalin, and Chiang Kai-Shek stated that they hoped that Japan would agree to surrender unconditionally and prevent devastation of the Japanese homeland and that they patiently awaited Japan's answer.

Reporters in Tokyo questioned Japanese Premier Kantaro Suzuki about his government's reaction to the Potsdam Declaration. Since no formal decision had been reached at the time, Suzuki, falling back on the politician's old standby answer to reporters, replied that he was withholding comment. He used the Japanese word *mokusatsu*, derived from the word for "silence." As can be seen from the dictionary entry quoted at the beginning of this essay, however, the word has other meanings quite different from that intended by Suzuki. Alas, international news agencies saw fit to tell the world that in the eyes of the Japanese government the ultimatum was "not worthy of comment." U. S. officials, angered by the tone of Suzuki's statement and obviously seeing it as another typical example of the fanatical Banzai and Kamikaze spirit, decided on stern measures. Within ten days the decision was made to drop the atomic bomb, the bomb was dropped, and Hiroshima was leveled. [1]

Almost without exception, whenever this story is told, mention is made of the poor translation job. One short magazine article [2] calls it "The World's Most Tragic Translation," "the ill-chosen translation of a common Japanese word," "disastrous oversight in this most important of all messages," and "that inauspicious translation." Indeed, there seems to be little question about the translator's culpability.

Many people, especially non-linguists, seem to feel that every word in one language has an exact counterpart, a perfectly equivalent match, in every other language. So, given a word in Language A, this can mean only one thing in Language B, and that one thing will be exactly what was meant in Language A. Obviously, this is not true. Different peoples with different cultural backgrounds view things differently and their languages reflect this difference of viewpoints. For example, given six animals common to several regions, one language group may categorize the animals into two classes because of size and have only two words in their language (one for large animals, one for small animals); another people may use the animals' eating habits as their criterion and also have two words (one for carnivores, one for herbivores) encompassing different groups of animals. Another people may subdivide the animals by color and end up with four words in their language, and yet another people may not do any subdividing at all; so they'll have six words to use in talking about these animals. It's also possible that some other group may have separate names for the male and female of the species (as English does for *ram* and *ewe*, *gander* and *goose*, etc.), so that they will have twelve different words! In addition, there are other criteria that could be used, and the number of words could be increased or decreased; or several languages might have the same number of words but, using completely different criteria for their subdivision, they would concern completely different things.

Another linguistic problem that keeps every word in one language from having a counterpart in every other language is that often something which is commonplace to speakers of one language will be totally unknown to speakers of another tongue. They have no concept of the thing; so how can they have a word for it? This is a problem frequently encountered by Bible translators. How, for example, do you translate "Lamb of God" into an Eskimo dialect whose speakers don't have the slightest idea of what sheep are? Or "anchor" into a language spoken by nomadic dwellers in the midst of the Sahara Desert?

Quite often a term *can* be translated word for word, but the resulting translation carries completely disparate meanings in both languages because of cultural differences. A missionary in Africa

ran into this problem with the sentence, "Behold, I stand at the door and knock" (1 John, 3:20). In that region only a thief knocks on doors; if anyone responds, the intruder runs away. Visitors with honest intent shout the name of the person in the hut. The translator solved the problem by rendering the verse, "Behold, I stand at the door and call." [3]

Bible translation points up some of the problems that exist for all translators, and recent translations of the Scriptures into English show an acute awareness of the specific problem that not every text in one language can be readily, idiomatically, and unequivocally translated into another tongue. Thus, while the King James, Douai, and 1917 Jewish Publication Society versions of the Bible are singularly free of—or, at least, mighty sparing with—notes, more recent translations—be they Protestant, Catholic, Jewish, non-denominational, or interdenominational (such as the 38-volume Anchor Bible)—contain copious notes, explanations of doubtful readings, alternate versions, comments on uncertainty about certain obscure words, etc. It came as a shock to many people to discover that Moses and Jeremiah and Jesus and Paul didn't speak the majestic-sounding English prose of the King James version as their native tongue, but it comes as an even greater shock to find out that the Bible they grew up with may have contained some questionable translations.

This really shouldn't be too surprising, though. Besides the cultural differences existing between the original writers of the Bible and us, there is also the time difference, and modern Bible translators are quite correct in admitting their inability to capture the precise shade of meaning in an old or obscure Hebrew or Greek word. Their willingness to admit their limitations should serve as a shining example to all translators to be honest with the people who will have to read their translation. If a word is capable of variant translations, this fact should be conveyed to the reader, provided that there is no way of clearing up the ambiguity by research.

Whoever it was who decided to translate *mokusatsu* by the one meaning (even though that is the first definition in the dictionary) and didn't add a note that the word might also mean nothing stronger than "to withhold comment" did a horrible disservice to the people who read his translation, people who knew no Japanese, people who would probably never see the original Japanese text and who would never know that there was an ambiguous word used. As a matter of principle, that unknown translator should have pointed out that word has two meanings, thereby enabling others to decide on a suitable course of action.

But that choice was lacking; the harsh decision was made on the basis of the one translation available; and the consequence was disastrous. To be sure, the case of *mokusatsu* is an extreme example, but there are many, many other words in scores of other languages that have equally discrete nuances of meaning. Choosing the wrong meaning or failing to inform a decision-maker that there are such nuances could well have unexpected repercussions. Suppose that someone quoted a report about a Cuban farmer who placed a bomb on his farm near the fences of the U. S. Naval Base at Guantanamo. Can you imagine the reaction of the Joint Chiefs of Staff, especially if no one told them that the Spanish words translated as "placed a bomb" might also mean "installed a pump"?

Many translators hesitate to admit publicly that they can't give a precise equivalent for an ambiguous text. They feel that saying that a word can have two or more equally possible meanings is a sign of their inadequacy. They fail to realize that if they pick the wrong one, and later results prove that they guessed wrong, their prestige will sink even lower than if they had insisted that the term was ambiguous. But the fault isn't always the translator's. Quite often non-linguist supervisors fail to understand that languages don't always have a one-to-one correspondence in their vocabularies and they will insist that the translator give them just one meaning and "cut out that 'ambiguous word' nonsense!"

Even if some high American officials ever knew that the Japanese Premier had used the word *mokusatsu*, they probably wouldn't believe that it could be translated in either of two ways. More than likely some high-ranking officer (probably a bird colonel) in the Pentagon asked the top Japanese translator (probably a nisei Pfc) about *mokusatsu* and then refused to believe the story about two meanings. You can almost picture this colonel pounding his desk and red-facedly bellowing, "Whaddyamean telling me that the word means either 'I am maintaining silence' or 'I'm treating it with contempt'? Dammit, private, I can't go tell the Joint Chiefs something like that! I gotta give 'em straight facts, not multiple-choice tests! Now you quit beatin' around the bush and give me one answer that I can tell 'em!"

But even if this scene never happened in the Pentagon, the fault for the *mokusatsu* incident is not entirely the translator's. Believe it or not, the real culprit is no less a personage than Kantaro Suzuki, the Japanese Prime Minister himself! [4] After all, there would have been no translation problem if he had not used an ambiguous word for such an important statement.

However, politicians are notorious for preferring words that are either meaningless or so full of meanings that no one can be sure of

just what they do mean. In all probability the word *mokusatsu* was well beloved by Japanese officials as their equivalent of "No comment!" simply because it does have such a broad spectrum of meanings. A politician could use it and not really be saying anything he couldn't squirm out of later; but it also left him the opportunity to claim later that he had long been against the course of action under discussion. Terms like these are the ones that Theodore Roosevelt once referred to as "weasel words" because politicians have sucked all the meaning out of them the way that a weasel sucks out the contents of an egg.

Politicians aren't the only ones who use meaningless words. Newsmen are also guilty of this sin, and so are many people right here in our own agency. For instance, TV journalists and NSA writers both seem to like the word "indicate" and they use it over and over and over in a variety of senses ranging from "hint" to "show conclusively," and even as a synonym for "say." During World War II the word "aircraft" became the "in" word, despite its vagueness. In the interim since it gained such popularity, I have seen the word used to translate foreign texts where the original was extremely specific in talking about one airplane, several airplanes, a bomber, crop-dusters, several helicopters, a trainer, a seaplane, or a DC-4, among others. But the translators or reporters uniformly used the word "aircraft" in titles and texts. I have seen items with the word "aircraft" used throughout, and it was necessary to read several paragraphs—in one case, three whole pages—to find a pronoun referring to the "aircraft" so that I knew whether the word was singular or plural. And I've even seen a few documents in which, even after careful reading and rereading, it was absolutely impossible to determine whether one or many "aircraft" was or were meant.

There are many other examples of ambiguous words that could be cited, but I think that these two exemplify the second lesson that can be learned from the *mokusatsu* incident: Try to avoid ambiguous words!

Some years ago I recall hearing a statement known as "Murphy's Law" which says that "If it can be misunderstood, it will be." *Mokusatsu* supplies adequate proof of that statement. After all, if Kantaro Suzuki had said something specific like "I will have a statement after the cabinet meeting," or "We have not reached any decision yet," he could have avoided the problem of how to translate the ambiguous word *mokusatsu* and the two horrible consequences of its inauspicious translation: the atomic bombs and this essay.

NOTES

- [1] The story is told in greater detail in William Craig's *The Fall of Japan*, published by Dial, 1967.
- [2] *The World's Most Tragic Translation*, unsigned article in *Quinto Lingo*, January 1968, p. 64.
- [3] The linguistic problems cited in the preceding paragraph were solved by equally ingenious methods. Many Eskimos catch a seal, cook it, and eat it at a communal meal before a big fishing expedition; so translators more or less equate this propitiatory seal with the sacrificial lamb eaten by the Israelites before their departure from Egypt (Exodus, 12: 3-11) and the expression "Seal of God" is used for "Lamb of God." At the other extreme, Saharan natives do use pegs or picket stakes driven in the sand as a means of "anchoring" their camels and horses at night. So the American Bible Society rendered "steadfast anchor for the soul" (Hebrews, 6:19) as "steadfast picket peg for the soul" when they translated the Bible into the language of one of the Saharan nomadic tribes.
- [4] Unfortunately, Mr. Suzuki will never know that I blamed him for all this trouble. He died 17 April 1948.

Machine Translation—Where Is It Today?

BY CLYDE D. McDONALD

INTRODUCTION

A standard textbook on data processing will probably list language translation as one of the uses of modern computing machinery, but will not go into any degree of detail. This is probably the easy way out for the author, for I submit that computer translation is probably the most controversial and misunderstood computer application yet encountered. During the past ten years, magnificent progress has been made in applying computers to business operations (e.g., banking), and in the five-year period preceding that, the computer had similarly revolutionized scientific computation. The number of computers in the United States has grown in exponential fashion, as has the investment in computer facilities. Breakthroughs in program compilation, computer speed, memory devices, and so on have cascaded in dizzying announcements. Only the sky seems the limit! But in computer translation, there is a disquieting note, as recorded in an editorial in the *Washington Post*, November 27, 1966: "If the computer has really failed in the relatively simple task of transferring thought content from one language to another, the result is to challenge the whole broad assumption that we are entering upon a new 'age of the computer'." What is this phenomenon being commented on here? Is the task really "relatively simple?" Indeed, it is not, for language translation falls into that annoying category of computer applications known as "ill-structured," a term which describes problems for which no straightforward computing procedure can be obtained from analysis. Another important point illustrated by the editorial mentioned above is that the press, and indeed the public, has been generally unkind to computer translation. A viewpoint has developed that language is one area that is human, not to be invaded by those dumb-brute machines which have taken over our payrolls, gasoline bills, and supermarket contests. Characteristic of this human pleasure at computer failures is the class of joke which has grown up in company with computer language translation; (sample: the computer translates the foreign language equivalent of "hydraulic ram" into English as "water goat").* Computer trans-

*See also Arthur Salemme, "The Case of the Volatile Mouses," *NSA Technical Journal*, Vol. V, Nos. 3 & 4, Fall, 1960.

lation is not alone in the group of "ill-structured" problems for which attempts at computerization have been made, but it does represent a problem which was recognized early and on which large sums were expended over a period of years. Thus it may be a test case for computerization of the "ill-structured" problem, since it is furthest evolved. It is a case to be considered from all sides in search of new meanings since it may apply to other "ill-structured" problems. With this background information I began my search for published material on automatic language translation. An attempt will be made in succeeding paragraphs to expose some of the more conclusive material reviewed in the research for this paper.

TRANSLATING SYSTEM PARAMETERS

In concept, a mechanical translation system would consist of three parts: an input, a processor, and an output. Output devices which would be acceptable are available, and input devices as well; the latter would include the class of optional character recognition equipment now feasible. The processor is the heart of the problem. For current purposes, a large-scale, general-purpose computer appears to be adequate. Thus, it appears that there are no significant equipment problems relating to computer translation, but rather, that such problems as exist lie in the translation procedures used in the programs. Five different approaches, of increasing complexity, are cited and defined as follows:

Item-for-Item Substitution—Linguistic items, whether they be parts of words, whole words, or groups of words, are processed independently and sequentially converted into corresponding items in another language. In this approach, for large dictionaries, clerical errors can be a problem, and the usual multiplicity of equivalents must be resolved in some manner. It has been found to be a useful machine-aid process for human translation.

Morphological Processing—This method is a refinement of the item-for-item process; it takes into account the rules for the creation of derived forms of words, in addition to the dictionary forms.

Syntactic Translation—This process takes into account the rules by which words are arranged into sentences. Automatic parsing, the analysis of the source language sentences by functional segment (as, subject, predicate, etc.), was unusually difficult, but after many years of effort reasonably complete programs for automatic parsing have been developed. However, the problem of constructing grammatical sentences in the object language has not been developed to the same degree.

Transformational Translation—Transformation rules indicate how various syntactic constructions in a language are related to, or may be derived from, other constructions regarded as simpler or more basic. Translation by transformational processing would involve the use of transformational grammars both for the analysis of the source language and for the generation of the object language. Fairly good transfor-

mational grammars for English have been developed but are incomplete for other languages.

Semantic Translation—Semantic translation would be based on meaning, for, after all, it is the meaning which must be conveyed. Research has been conducted in this area, but much remains to be done.

The current state of automatic language translation is such that, for some pairs of languages (for example, Russian to English), the procedure is conducted at a level somewhat below syntactic translation. This has been attained through the expenditure of significant amounts of time and money, and with resulting output which can at best be described as only fair. Human post-editing remains a requirement. (Mr. Richard See has provided information on a recent development in mechanical translation which I was not aware of at the time my paper was prepared. The Atomic Energy Commission is now using a mechanical translation program, developed at Georgetown University, at their Oak Ridge, Tennessee, facility. Rapid service is being provided to about twenty users of Russian translations, without post-editing. This was reported in *Science*, Vol. 155, January 1967, p. 59 and also by Francois Kertesz, "Babel—or Translation by Man and Machine," Oak Ridge Report ORNL-TM-1723.)

AN EXISTING SYSTEM

Most of the support for automatic language translation in the United States has been provided by the Federal government. During the ten-year period, 1956–1965, this support amounted to approximately \$20,000,000, about half of which was provided through the United States Air Force. The major Air Force effort has been the development of the AN/GSQ-16 (XW-2) Automatic Language Translator for the USAF Foreign Technology Division. In a report produced in 1965, the Arthur D. LITTLE Company has evaluated the Foreign Technology Division's system, which is described as a machine-aided system. The purpose of the LITTLE study was not to justify or to defend previous expenditures, but to consider the alternatives of continuing the system, improving it, or abandoning it. The analysis was conducted in terms of quality, cost, and timeliness. The evaluation concluded that the quality of output is equal to that of standard human translations. The system translation cost at the Foreign Technology Division was found to be about \$36 per thousand Russian words, excluding overhead and equipment amortization; in-house human translation cost, excluding overhead, was about \$40 per thousand Russian words. Contract translation cost was about \$33 per thousand Russian words, including contractor's overhead. Post-editing and recomposition were the main cost components in

the process, accounting for over 70% of the total; input processing accounted for only 11%. With regard to timeliness, and excluding high-priority items, all forms of human translation were found to require 85-110 days for translation of a moderate-sized article. In contrast, machine translations of articles of similar length required no more than an average of 45-50 days. The study reached the conclusion, for the near-term, that the system is both technically and economically feasible, and that it should continue to be operated as a production facility devoted to producing high-quality service to users of translations.

ECONOMICS OF TRANSLATION

In April, 1964, with funds made available by the Department of Defense, the National Science Foundation, and the Central Intelligence Agency, the Automatic Language Processing Advisory Committee of the National Academy of Sciences—National Research Council was organized under the chairmanship of Dr. John R. Pierce. The report of this committee, published in late 1966, may be regarded as a defense of human translation in the near future. "The Committee is puzzled by a rationale for spending substantial sums of money on the mechanization of a small and already economically-depressed industry with a full-time and part-time labor force of less than 5,000." The Committee's report is not optimistic with regard to future demand for larger quantities of translations, even if they were inexpensive and of high quality. A major criticism of current machine translation in the report is that post-editing takes almost as much human time as human translation from the original. The Committee believes that present practices in translation can be improved; some machine aids to translation, such as automatic dictionary lookup, can be used. Some technical articles of little value can simply remain untranslated. United States scientists should easily be able to learn enough Russian—the most important foreign technical language—to read papers in their field. The Committee's report, indeed, presents such a dreary outlook that "the chairman was asked to prepare a brief statement of the support needs for computational linguistics, as distinct from automatic language translation. This request was prompted by a fear that the Committee report, read in isolation, might result in termination of research support for computational linguistics as well as in the recommended reduction of support aimed at relatively short-term goals in translation." In the requested statement, Dr. Pierce said that "among the important kinds of research that need to be done and should be supported are (1) basic developmental research in computer methods for handling

language, as tools to help the linguistic scientist discover and state his generalizations, and as tools to help check proposed generalizations against data; and (2) developmental research in methods to allow linguistic scientists to use computers to state in detail the complex kinds of theories (for example, grammars and theories of meaning) they produce, so that the theories can be checked in detail.

"The most reasonable government source of support for research in computational linguistics is the National Science Foundation. How much support is needed? Some of the work must be done on a rather large scale, since small-scale experiments and work with miniature models of language have proved seriously deceptive in the past, and one can come to grips with real problems only above a certain scale of grammar size, dictionary size, and available corpus.

"We estimate that work on a reasonably large scale can be supported in one institution for \$600 to \$700 thousand a year. We believe that work on this scale would be justified at four or five centers. Thus, an annual expenditure of \$2.5 to \$3 million seems reasonable for research. This figure is not intended to include support of work aimed at immediate practical applications of one sort or another."

LINGUISTS AND PROGRAMMING LANGUAGES

As programming languages have become more sophisticated, the relation between them and linguistic science has become more recognizable. Before FORTRAN, there was little effect of linguistics on programming. In the period from FORTRAN to ALGOL, 1956-1960, the connection was almost exclusively in terminology; words and definitions, but not theory and technique, were borrowed from linguistics (for example, grammar, syntax). Intense borrowing has occurred during the period from ALGOL to the present from current mathematical linguistic theory, technique, and notation. It is now possible to teach syntax analysis of programming languages, and to construct programs which are general-purpose syntax analyzers in the sense that they parse any programming language of a given type. Programming has, in turn, had its effect on linguistics by providing linguistics with real models and by providing pressure on linguistic theory, directing it to problems particularly relevant to computation. Special programming languages (SNOBOL and COMIT) have been developed for solving certain linguistic problems. What of the future relation between programming and linguistics? According to the report of Dr. Pierce's Committee, "These (programming) languages will become sufficiently complex so that a theory of their semantics or evaluation will be a sufficiently interesting model for the equivalent problems arising in natural language translation."

"Similarly, there will be a reverse flow from the development of semantic theories within natural linguistics into mathematical linguistic models, which, in turn, will influence programming."

FOREIGN EFFORTS

In order to do work in machine translation, it is almost essential that computers be available to the researchers, and American efforts have been enhanced by the enormous expansion of computing facilities available in this country. The Russian effort is stronger in theory than in practice, being associated with the Soviet interest in cybernetics. They appear to be short on computer time, and deficient in alphameric input/output equipment. There is also a French effort, and a small but practical English one. West Germany is oriented toward machine aids to translation and has made some progress. There is some study in computer translation being conducted in Japan, and, presumably, in Communist China.

CONCLUSIONS

From a period of growing interest, publication, and expenditure in the automatic language translation field which ended about 1961, we have seen that a period of uncertainty, controversy, and evaluation has followed. Tentatively, we may mark the end of this period with the publication of the Pierce Committee report in November 1966. Perhaps only now can the relationship between computers and linguistics begin to show its true depth. Clearly the problems of computers in translation are more difficult than many realized, but the rewards, I believe, are potentially greater than imagined by most. Man (whatever his language, and whether he speaks or writes) must co-exist with the computer (whatever its language and regardless of the I/O devices). The pressures to communicate with the computer (or at least, to describe processes to it) in increasingly natural English will reinforce man's efforts to communicate, using the computer as an intermediary, with other men who speak a different language. The possibilities of real-time computer interpretation can add an entirely different perspective to the economic outlook for computer translation. Certainly language is an heuristic phenomenon for the human beings who created it and use it, and perhaps this is the approach that will ultimately be required. On the other hand, if it be granted that dictionaries and grammars have a stabilizing effect on language (especially if government sponsored), is it not possible that the future may bring with it increasing demands for rigidity in language? These forces of standardization can be seen at work in a related field—indexing and information retrieval. Faced with vexing di-

versity in terminology, one frequently has recourse to glossaries which define the "right" way to say a thing. In another field, the computer today is enforcing its style of writing on the computer programmer. Perhaps the future will bring a resurgence of the authoritarian grammar of the one-room school house. For those who might feel this approach to be too rigid for human acceptance, I wonder if they feel the same regarding the publication of a new glossary? Standardization of language may already be making inroads and may prove more tolerable in practice than in contemplation.

Acknowledgments

Mr. Richard See of the National Science Foundation was extremely helpful to me in providing guidance and the loan of documents; he provided the initial background information which directed and focused my reading. Dr. Sidney Jaffe (A7) and Mrs. Mary J. Fabik (P2), were helpful in providing documents for study. The five approaches to mechanical translation described under "Translating System Parameters" were derived from Mr. See's 1964 article in *Science*.

BIBLIOGRAPHY

- [1] Paul W. Howerton, "The Parameters of an Operational Machine Translation System," *Mechanical Translation*, Vol. 6, November 1961.
- [2] Anthony G. Oettinger, *Automatic Language Translation*, Harvard University Press, 1960.
- [3] D. Yu Panov, *Automatic Translation*, (translated from the Russian by R. Kisch, edited by A. J. Mitchell), Pergamon Press, New York, 1960.
- [4] John R. Pierce, *Languages & Machines, Committee Report*, National Academy of Sciences, Washington, 1966.
- [5] Richard See, "Mechanical Translation and Related Language Research," *Science*, Vol. 144, No. 3619, May 8, 1964.
- [6] Victor H. Yngve, "A Programming Language for Mechanical Translation," *Mechanical Translation*, Vol. 5, No. 1, July 1958.
- [7] *An Evaluation of Machine-aided Translation Activities of FTD*, Arthur D. Little, Inc., Contract AF33(657)-13616, 1965.
- [8] *Current Research & Development in Scientific Documentation*, National Science Foundation, 1966.

Employment of Military Linguists

BY JACOB GURIN

Some of the tricks which one's memory plays make the past seem so much more exciting and colorful than it really was. Embellishments and exaggerations slip unnoticed into the body of wartime stories as a result of passing years and numberless repetitions, and they serve to define and lend dimension to what otherwise might be a series of flat and colorless occurrences in the life of a G.I. But these are the recollections we have, and perhaps they may serve as a source of instruction, warning or amusement to others.

One of the unsung heroes of World War II was the linguist serving in Military or Naval Intelligence. My own experience was as a translator and interpreter of the Japanese language, one of a collection of odd birds who decided somehow that they could serve their country more effectively that way than in any other capacity.

In addition to all the other problems facing the U. S. at the outbreak of the war, we had the Japanese language. It seemed reasonable to expect that the Services would contain types who knew German or Italian, but where would we find anyone who knew Japanese? True, there were a few reliable sources of this skill. A number of missionaries had made Japan their life's work but had returned to the U. S. as conditions became more difficult for Americans there. A few academic types specialized in Japanese history, language and literature, and there was the handful of businessmen who had worked for years in Japan. But except for the Nisei (Americans whose parents were Japanese immigrants) the number of Americans who were conversant with Japanese was small.¹ The missionaries were seldom of military age, and the businessmen tended to run to specialized vocabularies and fluency in restricted areas of conversation or reading. Their children, however, provided the Services an aristocracy characterized by such feats as chattering in Japanese with amazing fluency about all sorts of trivia.

The Services had been training a small number of linguists all along as a matter of policy. It quickly became apparent, however,

¹That eminent authority on things linguistic, Life Magazine, in September 1942 estimated that fewer than 100 Caucasians in the U. S. were competent Japanese linguists.

that the number would have to be increased enormously if we would equip our side with the linguistic skills which the war required. The Navy opened a school at Boulder, Colorado, and later on another at Stillwater, Oklahoma. The Army's schools were at Ann Arbor, Michigan and Minneapolis, Minnesota.

Our teachers were Japanese, Nisei, and Korean. They displayed varying degrees of teaching skill, but were dedicated to their task. Few had experience as teachers, and fewer still had ever been called upon to teach the Japanese language. We all learned together.

We had both men and women teachers, and we mimicked their pronunciation and choice of words. Later on, in Japan, I found out how lucky I was to have studied under male teachers. Some of those who were not so fortunate were greeted with giggles as they solemnly pronounced their Japanese. Just as in English you would not regard as standard masculine speech such phrases as "just a *darling* shade of gray," or "simply *dying* to meet that precious girl," so in Japanese a man is regarded with suspicion if he uses feminine words or intonations. If the maid knocks on one's hotel room door and he answers "*haire*,"² she will come in and change the linen or whatever. But if a presumably male voice simpers "*O-haitte kudasai*,"³ she may not even enter the room!

The students were a mixed bag. We had a former lion tamer from the Ringling Bros. Circus, a small-time stand-up comic, the former coach of the Tokyo Eagles baseball team, some businessmen who had dealt with Japanese-speaking communities on the west coast and had become interested in the language, BIJ's⁴ who had difficulty in sitting down to learn the written language, and the majority—college types who felt that the study of Japanese would be their cup of tea.

Most of us started from scratch in the language. The objective of the course called for feats of memorization, covering huge amounts of vocabulary and several thousand ideographs, including a surprising number which had absolutely no application outside of Japanese mythology. Other unlikely goals were: fluency in the spoken language; appreciating Japanese culture; understanding the workings of the Japanese mind. What success we did achieve was attributable more to the dedication of students and faculty than to any virtues in the educational policies and procedures.

²Come in.

³Do me the honor of entering my unworthy room.

⁴Born in Japan. Sons of missionaries and businessmen.

Our instructors repeatedly assured us that eventually we would come to understand what we were doing, that explanations were difficult or impossible, and that by some mysterious alchemy we would in time become sufficiently adept to do the war effort some good. Oddly enough, they were at least partly right!

The weakest part of our course, as we soon found out, was the attempt to teach us the spoken language. Theoretically we were provided the opportunity to practice conversation by teaming up several times a week with a fellow student, a Nisei who had a reasonable command of ordinary spoken Japanese. But his obvious boredom and resentment quickly shattered our pathetic attempts to converse, and after a while the whole thing was given up as an unworkable idea. We did not fare any better in understanding the Japanese movies which were used as teaching aids. During the showing of one of these, a domestic drama of some sort, the students cheered mightily when, after about one hour of total lack of comprehension, they understood the maid when she knocked on the door and said "Excuse me!"

After a year of training, we were declared ready to go into operation. Our program had been amazingly comprehensive, and even included elementary training in judo. Because of my large size, I was always used to illustrate the advantages of judo in permitting a small man to deal effectively with a large one. In damp weather I still ache as a result of some of those falls. We also practiced bowing and squatting on our heels for prolonged periods, since we understood that these were favorite pastimes in Japan. Having added to this list of accomplishments mastery of the words and music of "Kimigayo," the Japanese national anthem, a pocketful of Kanji cards⁵, and a box of dictionaries, we went out into the world convinced that what we really needed was another year of training!

We linguists were a scarce resource. Our paltry number was splintered in the attempt to cope with the many and pressing needs for our services. Some of us went to the Central Pacific, some to the South Pacific, some to the Southwest Pacific, and some even found their way to Washington and the Navy and Army predecessors of NSA.

It was my good fortune to be sent to the Southwest Pacific where, shortly after my arrival, I was given my first document to translate. It was a diary taken from the breast pocket of the uniform of a dead Japanese soldier. The book measured about 1½ by 2½ inches, and although the entries had been made in a fine hand, they proved almost undecipherable as far as I was concerned. Could it have been

⁵Small cards with Japanese ideographs on one side, and pronunciation and meaning on the other—indispensable to the student.

the bullet hole through the center of the book? Or the fact that the book and its owner had lain on the beach for several days half-in and half-out of the water, causing the ink to run? They had not taught us to cope with this at school! I saved face the next day, when I plowed triumphantly through a printed manual entitled "How to Dig a Hole."

I became a member of a unique organization entitled Allied Translator and Interpreter Service, better known as ATIS. Its headquarters were hidden away in an obscure suburb of Brisbane, Australia, called Indoropilly. The record of this remarkable outfit is quite impressive. In its few years of operation 14,000 prisoners were interrogated and two million documents were scanned, the more important ones being translated in whole or in part. ATIS was truly an international affair, encompassing the American Army, Navy, and Air Corps, the Australian Army, Navy, and RAAF, the New Zealand Navy, the British Army, Navy, and RAF, and featuring several Chinese generals.⁶

One of our big problems was to convince combat troops of the value of prisoners and documents. We had to argue convincingly, for our troops were understandably loath to take chances with an enemy who appeared to be surrendering but who would fling a grenade at the very last moment. The record shows a terribly small percentage of prisoners captured in relation to the number of dead and wounded. For example, at Guadalcanal 40,000 were killed and about 600 taken prisoner. At Buna there were 43,000 counted dead and only 625 prisoners, and this ratio remained fairly constant until just before the conclusion of the Philippine Campaign.

Once a prisoner was captured, he talked with entire freedom and without reserve. With only very few exceptions, mostly among naval prisoners, he displayed a complete lack of security training. The Japanese government boasted that no Japanese would permit himself to be taken prisoner, and so provided no instruction in how to behave if captured. Since the prisoners felt that capture was an unspeakable disgrace, and that knowledge of their capture would bring ruin and probable torture to their own relatives, every one, immediately upon capture, gave false information as to name and organization. Of course, one of our jobs was to trip up the prisoner in the fabric of his own lies, and then he invariably broke down and apologized. From that time on we would have his cooperation, and like most prisoners he would talk about going to the U. S., since he could never go back to Japan because of his disgrace.

Col. Signey Mashbir, who was in command of ATIS, tells the story

⁶I never did determine their precise function, but we were proud to have them to show off to visitors.

of a bearded prisoner who would not cooperate. Someone thought of threatening him with a shave, to be followed by a photograph of his true features and distribution of the picture in Allied and neutral countries as our first prisoner. When he learned of these plans he talked without reservation.

We had one prisoner whom we interrogated for 18 months. He had been a patent attorney in private life, and had a remarkable memory. Perhaps he did make up some of the details, but he supplied an incredible amount of valid information on a variety of technical subjects. He was a pet of ours, and we took him to the GI movies on the base and occasionally into town, where in his borrowed Yank uniform he looked like one of our boys. It was a good thing that the townspeople didn't know who he really was!

Sometimes our interrogations were not particularly fruitful, in the sense of bringing victory to our forces. One member of our staff, a white-haired ex-missionary, worked over the same prisoner for a week, and we all grew curious about the report which would result from this exercise. When it appeared at last it was almost entirely a conversion progress report, with the observation that, in time, the prisoner "could make a good Christian."

On another occasion, an American Naval officer, an Annapolis man, was privileged to grill a captured Japanese Naval officer, one of the few we ever got. It turned out that they spent a considerable portion of the "interrogation" comparing promotion dates for their classes and the careers of their classmates.

Sometimes we acquired prisoners in bizarre ways. After the Japanese had been defeated in New Guinea, they fled to the jungle, where food was scarce and the natives unfriendly. Occasionally a half-starved Japanese would be caught trying to look inconspicuous as he awaited his turn in a GI chowline, mess-kit in hand. Our outdoor movies, located at the edge of the jungle, also acted as a magnet, and you could be sure that we often had an unseen and silent audience, especially for any picture featuring Betty Grable, a special favorite among the Japanese military. For Betty they would creep up to the very edge of the clearing which served as our theater.

While the troops were requested to send in all enemy documents, flags, notebooks—anything, in effect, which contained Japanese writing⁷—censorship of outgoing mail provided many valuable items. One notable example was on its way from a soldier to his girl in Iowa and bore the comment: "How do you like the Jap writing? It looks like hen tracks, don't it?" These "hen tracks" were an Intelligence

⁷We promised faithfully to restore these to the finder after the war.

Report from Imperial Headquarters giving an exact record of Allied planes which had been shot down during the preceding month.

The value of these documents and the possible penalty for failing to turn them in was illustrated by the tragic story of a young lieutenant killed by machine gun fire from an emplacement he was leading his patrol to discover. In his musette bag was a document he had "souvenired" and kept six days earlier instead of turning it over to us for translation. It showed the exact location of the very gun that killed him.

Every scrap was treasured. Many were extremely messy, as could be expected, since we were not above working with items covered with blood, body fat, etc. ATIS developed a Document Restoration Section to restore charred and soiled documents for the benefit of the translator's eyes. Along with chemical baths and ultraviolet light to make illegible documents readable, we also set up a simple process of sponging and ironing pages of all documents on which writing was decipherable. Employing five WACs who had been laundry workers, an electric ironer and some hand irons, we processed about 20,000 pages a day.

As is true in most other professions, translating documents can become a routine occupation with only an occasional event worthy of recall. One such occurrence was related by A. W. (Mike) Foley, who had been doing a translating job for the Marines. He had to go through a warehouse full of captured ordnance, reading the labels so that the equipment could be sorted and some of it used against its former owners. It was a hot, dirty job, and when it was over the grateful Marines offered to share their chilled beverages with Mike. Although everyone else seemed to enjoy it, he thought that the drink had a strange taste, and on examining the label he found that it read FU-RI-T-TO (Flit).⁸

Our work as linguists would have been very much more difficult if we had not had the talents of the Nisei to support us. Those Nisei who worked at ATIS had at least a fair knowledge of Japanese, and some were truly outstanding in both Japanese and English. Our best teams consisted of a first-class Japanese linguist, usually a Kibei (American Nisei who had received his schooling in Japan) and another man, usually Caucasian, whose English was good and Japanese at least fair. For highly technical subjects we would add subject-matter experts to the team.

Life for a Nisei in the Army could be dangerous. Those serving with combat troops were under suspicion from their Caucasian comrades and at times were fired at from both sides. The Australians were at

first loath to accept their aid, but there simply weren't enough Caucasians for all the urgent language tasks. It was found to be necessary to issue to each Nisei serving with Aussies an identity card showing him to be a member of the Australian Expeditionary Forces. They were urged to wear "digger" hats and we tried to make sure that they were always accompanied by Aussie officers. The Nisei soon proved their worth, and all who served with Australian combat troops received letters of commendation.

While most of the Nisei remained enlisted men, regardless of their obvious merits as linguists and intelligence specialists and their skill as soldiers, just before hostilities ceased permission arrived to commission one hundred of them. Many of these were my good friends. One in particular was George Kayano, who wrote the following about his experiences:

"Yesterday I was commissioned 2nd Lt in the Army of the U. S. Sure makes one feel odd. For a while I felt lost.

Today I received my first assignment. I am to be one of the few to witness the arrival of the Emissaries from Japan, to receive the surrender instructions from the Supreme Allied Commander of the Pacific. I was worried about the coming great event, also realizing that the whole world was waiting for this moment."

He then described in detail exactly what happened, the preparations made to receive the Emissaries, the landing of the plane, the debarking of the Japanese mission. He recorded his conversation with the Colonel who rode in his car, who stated that Hiroshima was 60 percent destroyed and that about 130,000 people had been killed or injured by the first atomic bomb. He concluded his letter with this paragraph:

"That was one night in my life that I shall never forget. I believe that the United States is the only nation where a nobody like me can participate in an international function."

Now the war was over, but there was as much need for our services as before, perhaps more. Senior linguists were promptly distributed among general and flag officers to act as interpreters. I drew the CG of 8th Army, Gen. Eichelberger, but envied my friend Major Bowers, who outranked me and therefore became Gen. MacArthur's interpreter. He was the one who got the glamour duties, such as greeting the Emperor and conducting him to the General's apartment on the occasion of their historic meeting.⁹

When we flew into Japan in the middle of August 1945 we didn't know what to expect. Neither did the Japanese, I suppose, for while several of us stood there, practicing our Japanese, we overheard one

⁸For you youngsters: *everyone* in 1945 knew that Flit was an insecticide.

⁹After that, his principal linguistic duties consisted of taking Mrs. MacArthur shopping in the Japanese market.

old lady say to another: "You see, I told you that English wouldn't be difficult. I can almost understand them now!"

Headquarters for the occupation was set up initially in Yokohama. It was an exciting place to be and an exciting time; one was conscious of participating, at least peripherally, in the making of history. We were surrounded by famous names: Wainwright, MacArthur, Stassen.¹⁰ One officer came up to me as I sat at the desk marked "G-2, 8th Army," and asked for an interpreter. He was hatless and wore two stars on the collar of his shirt; I didn't recognize him as any of our generals. He explained that he wanted to survey the damage done by the atomic bomb in Hiroshima and Nagasaki, and then I realized that this "general" was Adm. Richard E. Byrd, the famous explorer! For a moment dreams of glory flashed through my mind, but then I thought of the look he would have on his face as I thumbed clumsily through my dictionary looking up the Japanese for "hemoglobin" or "skin eruptions." I sighed and assigned him one of our better Nisei sergeants.

A little while later, while at the same desk, I got an urgent call from CIC—Tojo had shot himself and appeared to be dying. A linguist was desperately needed to get his last word. I reflected briefly on the prospect: There's Tojo, surrounded by reporters. He mumbles something and dies. "What did he say?" the reporters asked me. "I don't know, I didn't hear it clearly," I respond. Heretofore my interpreting experience had permitted me to say "Please speak more slowly. Please repeat that word. Say it another way. Please write it in Japanese." I found another good Nisei sergeant to attend to Tojo, admitting shyly that I would rather do the job myself but that I had to stay with the desk.

My boss, General Eichelberger, called me in a few days later and pointed to a samurai sword on his desk. It was an elaborate affair which had been in Tojo's family for about 800 years, and the scabbard was adorned with beautiful designs and valuable stones. The general looked at me rather sternly¹¹ and asked what the inscription said. It consisted of three ideographs, so elaborately engraved that I couldn't possibly know what they were. And even if I could make them out, nothing in my background or reading would have acquainted me with a Sino—Japanese aphorism of the 12th Century. I gulped and asked permission to study the problem, which meant

¹⁰Yes, CDR Harold Stassen, USN. I don't know why he was there, but he got a lot of attention.

¹¹Could it be that he was beginning to doubt the Army's wisdom in assigning me to be his interpreter?

that I sent in a Nisei sergeant who was well versed in classical literature. I half-hoped he would fumble it, but he had no trouble at all.

I finally found a job which suited my talents. I was put in charge of a train headed for western Japan to pick up liberated Allied prisoners of war. My first task after arriving in Osaka was to establish communications with Hqs back in Yokohama. I got on the telephone and found to my dismay that the operator couldn't make out my schoolboy Japanese and that I didn't have the faintest notion of what she was saying. I was rescued by an American Army lieutenant, a POW who had released himself. He asked whether he could help and to my surprise he did very well in spite of shocking grammatical errors. He explained that he had learned Japanese quickly because in camp one obeyed a barked order promptly or was encouraged by a rifle butt. On the way back to Yokohama with the trainload of liberated prisoners, I met a U. S. Navy lieutenant who had been trained as a Japanese linguist but who had carefully hidden this fact during his imprisonment for fear of reprisal.

The time had come for the Japanese to disband their Army. This task fell to the Imperial General Staff, and I was assigned to supervise the planning action. The Staff had prepared a list of military units, showing their locations and strengths, and had brought along an interpreter to read the list aloud in English. He was a Hawaiian who had gone to Japan as a youngster, had stayed on and became a lieutenant in the Japanese Army. His rendition of the list was hesitant and not entirely reliable, so I took over. As a linguistic challenge this was pathetic, but my performance caused a halt in the proceedings while the entire General Staff inquired of me how I had learned the language and shook their heads sadly when I told them about the service schools. They were astounded when I revealed to them that these schools had turned out hundreds of graduates who had mastered¹² their language. They appeared crushed at the thought that their language had offered them so little protection.

Communication between the occupation forces and the people often broke down in spite of good-natured attempts on both sides to make do with sign language and pidgin English. On one occasion the MP's were enforcing an occupation edict that the streetcars were not to be dangerously overcrowded. Several would-be passengers were chased off their precarious perches on one of the cars. One of this number persisted in going back to the car in spite of the MP's admonitions, and he finally had to be hauled off to the local police sta-

¹²A little boastful inaccuracy should be expected from a member of the victorious occupying forces.

tion. Luckily, one of our linguists happened to be at the police station¹³ and he informed the MP's that the man they had arrested was the ticket-taker.

Obviously, if one relates his wartime exploits, he saves the best for last. The final anecdote should permit the storyteller to retire in quiet triumph, leaving his audience to glow for a few minutes in admiration and envy. I had been asked by the War Department to recruit civilian linguists in the U. S. to replace the military who were leaving the service. My wife and I traveled about the country for several months, looking up old friends and subjecting them to merciless examinations for pitiful C.A.F. ratings. In Los Angeles we attended a dinner given by the Japanese-American Citizens' League, and my wife was the only one there who did not speak Japanese. I had taught her a few useful words and phrases, however, assuring her that with these she could sail through any emergency. At one point in the evening she turned to the young lady who was sitting next to her, and asked "Benjo-wa, doko desu-ka?" The response was an utterly blank look, and the question had to be repeated. Finally light dawned, and our young neighbor said, rather archly, "I know what you mean, but we *never* say it that way!"¹⁴ My wife informed me in a soft voice that it was Berlitz for her from that point on.

¹³For obvious reasons I won't identify him.

¹⁴It was as though she had asked her hostess, at a posh party, where the john was!

The Voynich Manuscript

"The Most Mysterious Manuscript in the World"

BY BRIGADIER JOHN H. TILTMAN

The following paper is a slightly expanded version of a paper which I delivered to the Baltimore Bibliophiles on March 4, 1967. I am fully aware of the inadequacy of my treatment of the subject. The paper is intended only as an introduction to the study of the manuscript for anyone approaching it for the first time.

INTRODUCTION

The Voynich Manuscript is a vellum book of over 200 pages. There is text on almost every page in an unknown script. There are also coloured drawings on all but about 20 pages.

Plate 1 will give you an idea of what the script looks like. *Plate 2* is an example of an illustrated page.*

To the best of my knowledge there is no confirmed solution of the script or any part of it, and the authorship and general dating of the manuscript is totally unknown. With the exception of a small number of later additions (not in the unknown script), the character of the script and general behavior of the symbols appear to be constant throughout the book. Opinions differ as to whether the whole is by one hand. There appear to be no erasures or corrections, which suggests that the manuscript as we see it is likely to be a copy of an original which may be of an earlier date. In any case the writing of the manuscript and the painting of the illustrations must have been a major undertaking. The late Father Petersen made his own transcription of the manuscript without the illustrations, and it occupied him for, I believe, three or four years.

Only a comparatively small part of this paper is original, i.e., has not appeared in print before. My purpose in writing it is to widen the circle of those who might be interested in the manuscript. I am a working man and have not been able to devote much time to its study and am fully aware of the many deficiencies in my knowledge.

DESCRIPTION AND HISTORY

In 1912 the manuscript was purchased by the late Mr. Wilfred M. Voynich (later a rare book dealer in New York) who "discovered it in

*All plates appear at the end of the article.

a chest in an ancient castle in Southern Europe." It is now owned by Mr. H. P. Kraus, the New York antiquarian bookseller, who has revealed that it was found at Mondragone. This is a villa in Frascati near Rome, built by Cardinal Altemps about 1570. In 1582 Pope Gregory XIII issued from Mondragone the bull reforming the calendar. The villa apparently continued in the Altemps family as in 1620 a later member of the family bequeathed the Mondragone library to the Vatican library. In 1865 the villa became a Jesuit college which was finally closed in 1953. From 1912 to 1919, Voynich attempted to interest scholars in Europe and America in solving the script, while trying himself to determine the origin of the manuscript.

The manuscript, when discovered, was accompanied by a letter, shown in *Plate 3*. With Mr. Kraus's permission I quote from his catalogue number 100, entitled *Thirty Five Manuscripts*, the passages "History of the Manuscript" and "Conjectures concerning the early history of the Manuscript:"

HISTORY OF THE MANUSCRIPT

The manuscript enters recorded history on the 19th of August, 1666, when Joannes Marcus Marci of Cronland sent the codex from Prague to Athanasius Kircher, at Rome, with a signed autograph letter, which is found loosely laid into the manuscript. It reads as follows (transl. from the Latin):

"Reverend and Distinguished Sir, Father in Christ:

"This book, bequeathed to me by an intimate friend, I destined for you, my very dear Athanasius, as soon as it came into my possession, for I was convinced that it could be read by no one except yourself.

"The former owner of this book asked your opinion by letter, copying and sending you a portion of the book from which he believed you would be able to read the remainder, but he at that time refused to send the book itself. To its deciphering he devoted unflinching toil, as is apparent from attempts of his which I send you herewith, and he relinquished hope only with his life. But his toil was in vain, for such Sphinxes as these obey no one but their master, Kircher. Accept now this token, such as it is and long overdue though it be, of my affection for you, and burst through its bars, if there are any, with your wonted success.

"Dr. Raphael, a tutor in the Bohemian language to Ferdinand III, then King of Bohemia, told me the said book belonged to the Emperor Rudolph and that he presented to the bearer who brought him the book 600 ducats. He believed the author was Roger Bacon, the Englishman. On this point I suspend judgement; it is your place to define for us what view we should take thereon, to whose favor and kindness I unreservedly commit myself and remain,

At the command of your Reverence
Joannes Marcus Marci
Of Cronland

Prague, 19th August, 1666."

The last numeral of the date has been altered by pen from "5" to "6," obviously by Marcus himself. The emperor Rudolph II (1552-1612) was a scholar rather than a man of affairs; he neglected his duties as ruler of his realm in order to devote

himself to the study of alchemy, astrology, and astronomy, and he was the patron of Tycho Brahe, Kepler, John Dee, and a host of other scientists and pseudo-scientists. He resided in Prague throughout most of his reign, where he assembled a great collection of books and art objects. Rudolf, after acquiring the book, apparently loaned or gave it to Jacobus Horcicky de Tepenecz (died 1622), whose name, "Jacobi à Tepenece," is written on the recto of the first leaf (erased, but easily visible under ultra-violet light). The form of the name shows that the book must have been in his hands after 1608, when the "de Tepenecz" was acquired by patent of nobility from the Emperor. Tepenecz was the director of Rudolf's alchemical laboratory and his botanical gardens.

Dr. Raphael (1580-1644), a lawyer and minor poet, who supplied information concerning the book to Marcus, was connected with the Imperial court under Emperors Rudolf II, Ferdinand II, and Ferdinand III, and thus was in an excellent position to have obtained knowledge concerning the codex. He was attorney-general of Bohemia under Ferdinand III.

Joannes Marcus (1595-1667), the writer of the letter, was Rector of the University of Prague, and a noted physician, mathematician, and orientalist. He was official physician to the Emperor Ferdinand III. In 1667 he was elected a corresponding member of the British Royal Society. He had studied under Kircher, at Rome.

The connection of the cipher manuscript with the famous Athanasius Kircher, S.J. (1601-1680), is especially intriguing. He was one of the foremost scholars of the Jesuit order, keenly interested in problems of decipherment and the author of three works on an attempted solution of the Egyptian hieroglyphics, and of another work (*Polygraphia*, 1663), on codes and ciphers in general. In the XXth century the codex was acquired by Wilfried M. Voynich, a dealer in manuscripts, in 1912, who discovered it in a chest "in an ancient castle in Southern Europe."

CONJECTURES CONCERNING THE EARLY HISTORY OF THE MANUSCRIPT

In his letter, Marcus says, "He (apparently the Emperor Rudolf) believed the author was Roger Bacon, the Englishman." Bacon (1214?-1294), famous scientist and philosopher of the Middle Ages, studied and taught at Oxford and Paris, and is believed to have died at Oxford. Professor Newbold, who considered this manuscript to be in the autograph handwriting of Roger Bacon, conjectured that his papers were acquired after his death by some English monastery; that the present one, on the dissolution of the monasteries after 1538, was acquired by John Dudley, Duke of Northumberland (1502?-1553); that John Dee (1527-1608) acquired it from the Duke or some other member of the Dudley family; and that Dee, who lived in Prague for several years and was personally acquainted with Rudolf II, sold it to him.

An interesting point is that John Dee, while in Bohemia from 1585 to 1588, possessed

"a book . . . containing nothing but hieroglyphicks; which book his father bestowed much time upon, but I could not hear that he could make it out"

(Sir Thomas Browne, quoting Dr. Arthur Dee, 1579-1651, the son of John Dee.). It is indeed very probable that the present volume was the one which Arthur Dee saw in his father's hands; it would seem from the tenor of this reference, that the elder Dee was trying to extract some meaning from the book. It is quite possible that the Emperor, after purchasing it, had entrusted it to him for decipherment. The word "hieroglyphicks" would not, of course, refer at the time to Egyptian writing specifically, but to any secret alphabet such as that of the present codex.

Plate 4 is a photograph of the bibliographical description of the manuscript in the Kraus catalogue.

In the manuscript there are 125 pages of botanical drawings, 26 of astronomical (or astrological drawings), 28 of so-called biological drawings, and 34 of pharmaceutical drawings, nearly all accompanied by text. At the end there are about 20 pages of script without illustrations. *Plates 5 to 15* show examples of the appearance of the pages.

Plate 5.—The plant depicted here has been identified as some sort of Bindweed (*Convolvulus*).

Plate 6.—This seems to me a fairly natural representation of cross-leaved Heath (*Erica*).

Plate 7.—This has been identified as sunflower, giving rise to the hypothesis that the manuscript cannot be dated earlier than 1493, when Columbus introduced sunflowers into Europe. This identification has not been universally accepted.

Plate 8.—This is an example of the many drawings which appear to be composite and cannot be identified as any one plant. I should perhaps apologise for the lack of definition in some of these illustrations. This one was prepared from my photostat which was reproduced from Mr. Friedman's photostat, which was itself a copy of a photostat made many years ago by Father Petersen.

Plate 9.—This is one of the astronomical illustrations.

Plate 10.—This is one of the so-called zodiacal illustrations. There were presumably either one or two for each month, but the pages for two months seem to be missing. For each month there are 30 (or about 30) human figures round the edge of the circle.

Plate 11 shows two pages for April. One page has a black bull and 15 unclothed ladies, and the other, a white bull and 15 clothed ladies. All the zodiacal drawings carry the name of the month in the centre in a later hand and in readable script though the language has been disputed.

Plate 12.—This is one example of the illustrations in the biological part of the book. I have not myself studied these pages, and ideas as to their meaning advanced by specialists in medieval and early Renaissance history are completely outside my field.

Plate 13.—This drawing is an example of one of many pages which seem to comprise a sort of pharmacopeia.

Plate 14.—This is the first page. There is no title page, but this appears to be a table of contents, there being four paragraphs corresponding to the four illustrated portions of the manuscript.

Plate 15.—This is the final page and has on more than one occasion been assumed to be a key to a cipher used for producing the text as we

see it. You will see that only one short passage is in the unknown script used for the rest of the manuscript and that there appears to be some evidence of Latin words and at the end an unfinished Old English or Old German sentence.

PROPOSED SOLUTIONS

There have been three published solutions, none of them generally accepted.

(1) William R. Newbold, professor of moral and intellectual philosophy at the University of Pennsylvania, began work on the manuscript in 1919 and in April 1921 announced that he had discovered the key to a cipher, that he was convinced that the author was Roger Bacon, and that he had decyphered portions of it. He said that his decyphering proved that Roger Bacon had possessed both a telescope and a microscope, although history places the invention of these several centuries after his death, and further that one of the drawings depicted the great Andromeda Spiral Nebula, of whose existence he, Newbold, had been previously unaware. His solution, which was accepted at the time by Voynich and a number of scholars, was eventually demolished, particularly by Professor John M. Manly, Chairman of the English Department of the University of Chicago, in an article "Roger Bacon and the Voynich MS.," *Speculum*, July 1931, and now has no supporters. His complex method of decyphering (I quote from an article by Mrs. Friedman in the *Washington Post* of 5 August 1962) "was reducible to nine steps. The first and last of these, without any consideration of the intermediate abstruse and confusing processes, are utterly devoid of precision and are incapable of yielding one and only one plain language text—a rigid requirement of any legitimate cipher method. His first step was to convert the individual strokes of each symbol into Greek shorthand, a process of which Newbold himself said: 'I frequently find it impossible to read the same text in exactly the same way.' The reason for this, palaeographers say, is that what Newbold saw as separate strokes of a symbol are merely the results of the cracking, uneven spreading and fading of the ink, and the condition of the vellum because of the manuscript's age." Newbold died in 1926, and two years later his literary executor published a full-sized book from his voluminous notes and worksheets—*The Cipher of Roger Bacon*.

Plate 16 is an illustration from his book showing his interpretation in the form of shorthand strokes of a letter of the script; the page illustrated is one of the astronomical drawings and the arrangement of the stars near the portion of script which the author has chosen for illustration suggests Aldebaran and the Hyades.

The controversy over Newbold's solution left a legacy of ill-feeling which persisted for many years and which I found reflected in a letter which Charles Singer wrote to me in 1957.

(2) In 1943 a Rochester lawyer named Feely published a book entitled *Roger Bacon's Cipher; The Right Key Found*. Feely was the author of some published items in the field of the Shakespeare authorship controversy. His unmethodical method produced text in unacceptable mediaeval Latin, in unauthentic abbreviated forms.

(3) In 1946, a research scientist, Dr. Leonell C. Strong, published a different interpretation of part of the manuscript. He claimed to have decyphered part of two pages, one of them the Sunflower page shown in *Plate 7*. He said that his decyphrement revealed to him that the author was Anthony Askham and that the plain language text was in mediaeval English. Again his mediaeval English is not acceptable to scholars. He did not reveal the details of his key but his description of the cipher method makes no sense to cryptologists. However, in the process of preparing this paper I came across the following curious fact. Among the many editions of an illustrated English printed work known generally as *Banckes's Herbal*, the first dated 1525, are two usually attributed to Anthony Askham, dated 1550 and 1555. Their title pages read:—"A little herbal of the properties of herbs newly amended and corrected, with certain additions at the end of the book, declaring what herbs hath influence of certain stars and constellations, whereby may be chosen the best and most lucky times and days of their ministration, according to the moon being in the signs of heaven, the which is daily appointed in the Almanac, made and gathered in the year of our Lord God 1550, the 12 day of February by Anthony Askham Physician." The strange thing about this is that the astrological additions promised the reader, appear nowhere in either edition. Anthony Askham, a physician and clergyman, also published a number of almanacs. I managed to get a sight of one of these in a Library of Congress microfilm but found it only to be a Christian calendar on half a dozen pages.

In 1950 I was introduced to the manuscript by my friend, Mr. William F. Friedman, who gave me photostats of a few of the pages to work on, chiefly the unillustrated pages at the end. From these pages I made a preliminary analysis of the text, disregarding all but the most commonly occurring symbols. For purposes of the present paper, I have found it necessary to substitute for the symbols conventional and quite arbitrary letters and figures, as shown at the top of *Plate 17*, very similar to the system previously devised by Mr. Friedman for discussion of the problem with a study group which he brought together for a short time at the end of World War II. You will see that I have limited the number of symbols to 17. The second

symbols which I have placed at positions 1, 2, 15, 16 and 17, I have treated, rightly or wrongly, as variant forms. As a result of my analysis I made a report to Mr. Friedman in 1951 of which the following is the gist (only slightly revised).

"(a) Following are some notes on the common behaviour of some of the commonly occurring symbols. I would like to say that there is no statement of opinion below to which I cannot myself find plenty of contradiction. I am convinced that it is useless (as it is certainly discouraging) to take account at this stage of rare combinations of symbols. It is not even in every case possible to say what is a single symbol and what is not. For example, I am not completely satisfied that the commonly occurring A has not to be resolved into CI or possibly OI. I have found no punctuation at all.

(b) DZ and HZ appear to be infixes of D and H within T. The variant symbol represented by E appears most commonly at the end of a line, rarely elsewhere.

(c) Paragraphs nearly always begin with D or H, most commonly in the second variant forms, which also occur frequently in words in the top lines of paragraphs where there is some extra space.

(d) G occurs quite frequently as the initial symbol of a line followed immediately by a combination of symbols which seem to be happy without it in any part of a line away from the beginning. Otherwise it occurs chiefly before spaces very frequently preceded immediately by 8. Hence my belief that these two have some separative or conjunctive function. (I have to admit, however, that G also seems sometimes to take the place of O before D or H (though rarely, if ever, after 4); this is particularly noticeable in some of the captions to illustrations in the astronomical section of the manuscript—these most commonly begin OD or OH and it is here that we occasionally see GD or GH.)

(e) I have tried, for convenience of handling, to divide words into what I call "roots" and "suffixes." This arrangement is shown at the bottom of *Plate 17*. Regarding the second type of suffix, some of the combinations are so rare that I have been uncertain whether to take any account of them at all. Some are very common indeed. It seems to me that each of these combinations beginning A has its own characteristic frequency which it maintains in general throughout the manuscript and independent of context (except in cases where two or more A groups are together in series, as referred to later). These A groups, e.g., AR or AIIL, frequently occur attached directly to "roots," particularly OD, OH, 8 and 2. ODAIIL, 4ODAIIL and 8AIIL rank high among the commonest words in the manuscript.

(f) There are however many examples of 2, 3, 4 or even 5 A groups strung together on end with or without spaces between them. When this occurs, there appears to be some selective preference. For example, AR is very frequently doubled, i.e., AR AR, whereas AIIL which is generally significantly commoner, is rarely found doubled. Perhaps the commonest succession of three of these groups is AR AR AE. AE very frequently follows AR, but AR hardly ever follows AE.

(g) O, which has a very common and very definite function in "roots," seems to occur frequently in "suffixes" in rather similar usage to A, but nearly always as OR and OE. OR AIIL is very common.

(h) The behaviour of the A (and O) groups has suggested to me that they may in fact constitute some form of spelling. It might be, for instance, that the manuscript is intended to demonstrate some very primitive universal language and that the author was driven to spell out the ends of words in order to express the accident of an inflected language. If all the possible A and O combinations can occur, then there are 24 possibilities. They may, however, be modified or qualified in some way by the prefixed symbols D, H, OD, OH, T, S, 8, 2, etc., and I have not so far found it possible to draw a line anywhere. This, coupled with ignorance of the basic language, if any, makes it difficult to make any sort of attempt at solution, even assuming that there is spelling.

(i) E, usually preceded by A or O, is very commonly followed by D, much less commonly by H, with or without a space between. In this connection, I have become more and more inclined to believe that a space, though not intended to deceive, must not necessarily be regarded as a mark of division between two words or concepts.

(j) Speaking generally, each symbol behaves as if it had its own place in an "order of precedence" within words; some symbols such as O and G seem to be able to occupy two functionally different places.

(k) Some of the commoner words, e.g., ODCCG, ODCC8G, 4ODCC8G, ODAIIL, ODAAR, ODAE, 8AIIL, TC8G occur twice running, occasionally three times.

(l) I am unable to avoid the conclusion that the occurrence of the symbol C up to 3 times in one form of "suffix" and the symbol I up to 3 times in the other must have some systematic significance.

(m) Peter Long has suggested to me that the A groups might represent Roman numerals. Thus AIIL might be IIJ, and AR AR AE XXV, but this, if true, would only present one with a set of numbered categories which doesn't solve the problem. In any case, though it accounts for the properties of the commoner combinations, it produces many impossible ones.

(n) The next three plates show pages where the symbols occur singly, apparently in series, and not in their normal functions. The column of symbols at the left in *Plate 18* appears to show a repeating cycle of 6 or 7 symbols D (or H), O, 2, G, C, ?. In *Plate 19* the succession of symbols in the circles must surely have some significance. One circle has the same series of 17 symbols repeated 4 times. *Plate 20* also has an interesting column of symbols. In all three there are symbols which rarely, if ever, occur elsewhere.

(o) My analysis, I believe, shows that the text cannot be the result of substituting single symbols for letters in the natural order. Languages simply do not behave in this way. If the single words attached to stars in the astronomical drawings, for instance, are really, as they appear to be, captions expressing the names or qualities of those stars, there can hardly be any form of transposition system involved. And yet I am not aware of any long repetitions of more than 2 or 3 words in succession, as might be expected for instance in the text under the botanical drawings."

After reading my report, Mr. Friedman disclosed to me his belief that the basis of the script was a very primitive form of synthetic universal language such as was developed in the form of a philosophical classification of ideas by Bishop Wilkins in 1667 and Dalgarno a little later. It was clear that the productions of these two men were much too systematic, and anything of the kind would have been almost instantly recognisable. My analysis seemed to me to reveal a cumbersome mixture of different kinds of substitution. When I was attempting to trace back the idea of universal language, I came upon a printed book entitled *The Universal Character* by Cave Beck, London 1657 (also printed in French in the same year). Cave Beck was one of the original members of the British Royal Society and his system was certainly a cumbersome mixture.

Plate 21 shows the title page. His system is based on a dictionary code of about 4,000 words to which are allotted the numbers from 1-3999 in alphabetical order.

Plate 22 is a page from his dictionary covering the words "That (conjunction)" to "till or untill." When digital code-groups are unaccompanied by letters, they represent verbal infinitives, e.g., "to tickle"-3773. Groups are preceded by the letter R when a noun is indicated as in the case of most of the groups on this page; if the adjectival form is intended, the digital group is preceded by Q. In the case of words which the author regards as synonymous with words earlier in the alphabet, the earlier code-group is given, e.g., "to thinke"-1163, where "to cogitate" is found on an earlier page opposite 1163. Prepositions are represented by the corresponding word in Latin, e.g., "through"-*per*. Some very common words have trigraphic

equivalents beginning S and T to be found in a separate list of about 175 such words; these are repeated in their alphabetical place in the main dictionary, e.g., "that"-SNA. There are also on this page "thou"-E, "thine"-HE, and "this"-HO; also "Thursday"-+5 1484, i.e., "number five day." Plurals are designated by placing the letter S immediately after the digital code-group. There are also many other circumstances, particularly in the case of parts of verbs, which necessitate the prefixing of up to three letters before the digital code-group as the example taken from Cave Beck's book shows (*Plate 23*). As each word or concept treated as a word consists generally of a combination of letters and digits, it is necessary to put a comma after it to delimit it. I have wondered whether G and 8G in the Voynich Manuscript might be accounted for as representing "comma" and "plural comma." (I am aware that 8 occurs in other contexts in the manuscript, but so does S in Cave Beck's system.) Curiously, in his example (*Plate 23*) his substitution for "days" is given as 14848, not 1484S, as one would have expected.

Cave Beck starts his preface as follows: "This last century of years, much has been the discourse and expectation of learned men, concerning the finding out of an universal character" If this is to be taken literally, it carries the idea of universal language back to about the middle of the 16th Century. I tried in 1957 to trace back the idea of universal character, but I had little time to devote to this research and the earliest evidence I was able to find is contained in the following extracts from two lives of Bishop Bedell, who died in 1642, one by Burnet, dated 1692, and the other by Clogy, dated 1682, referring to a man named Johnston.

"But the Bishop finding the man had a very mercurial wit, and a great capacity, he resolved to set him to work, that so he might not be wholly useless to the Church; and therefore he proposed to him the composing of an universal character, that might be equally well understood by all nations: and he showed him, that since there was already an universal mathematical character, received for Arithmetick, Geometry and Astronomy, the other was not impossible to be done. Johnston undertook it readily and the Bishop drew for him a scheme of the whole work, which he brought to such perfection that . . . he put it under the Press, but the Rebellion prevented his finishing it."

"My Lord gave him a platform which he observed; all the difficulty was about the syncategoramata. He styled his books with spell. I heard that some part of it was printed; but the rebellion prevented finishing."

It would seem that these events cannot be dated later than 1641.

It is, however, difficult to date the manuscript anywhere near as late as this. Charles Singer, in a letter to me, put the date at late 16th Century. Professor Panoffsky and the keeper of the manuscripts at the Cambridge Library both independently insisted on a date within 20 years of 1500 A.D., and the manuscript as we have it may be a copy of a much earlier document.

HERBALS

At this point I propose to undertake a short digression on the subject of manuscript and early printed herbals. This may be thought irrelevant and my treatment will certainly be very superficial, but, if the plain text of the Voynich manuscript belongs to the illustrations on the same pages, as we have a right to expect in the complete absence of evidence to the contrary, then much the greater part of that text is related to plants. However, I have to admit that to the best of my knowledge no one has been able to find any point of connection with any other mediaeval manuscript or early printed book. This is all the stranger because the range of writing and illustration on the subject of the plant world from the early middle ages right through into the 16th and even 17th centuries is very limited indeed.

The first significant name is Krateuas, the Rhizotomist (rhizotomist meaning "root-digger") who was body physician to Mithridates VI Eupator, the King of Pontus who was defeated by Pompey and took his own life in 63 B.C. Krateuas wrote a herbal in which plants were not described but were depicted in figures which were followed by brief discussions of the medical uses of plants. Some of this herbal has been preserved by Dioscorides who was a physician attached to the Roman army in Asia about the middle of the first century A.D., and who is much the most famous name as a herbal writer in classical times. He wrote in Greek the work usually known as *De Materia Medica*. One of the most beautiful and, I imagine, the most valuable manuscripts in existence is the so-called Juliana Anicia codex at Vienna, written about 512 A.D. for presentation to Juliana Anicia, whose father was for a short time western Roman Emperor in 472 A.D. Two facsimiles of this manuscript have been published, the second in five parts, of which the first two are in the Garden Library at Dumbarton Oaks. Part of the manuscript is a text of Dioscorides, and the whole is magnificently illustrated in colour. I have reproduced here in *Plates 24 and 25* two pages of the manuscript. Charles Singer, in a most interesting article in the *British Journal of Hellenic Studies* for 1927, has restored about ten of the original drawings of Krateuas from it.

Plate 26 shows Krateuas engaged in painting a Mandrake held by Epinoia, the genius of Intelligence, while Dioscorides beyond writes in a book. This representation of mandrake as a nude figure, male or female, with leaves sprouting from or replacing the head, persists in illustrated herbals into the era of early printed books and even into the 17th century, carrying with it a collection of superstitions, some of them to do with the precautions to be taken when digging it up.

In 1957 I paid visits to a few specialists in early herbals in England. Among them I saw the late Dr. T. A. Sprague in Cheltenham and

showed him a few specimen photostats of herbal drawings from the Voynich manuscript, of which he had been previously unaware. As he looked at them he became more and more agitated and eventually said

"Do you know what you are asking me to do? I have spent the last twenty years of my life trying to identify the plant drawings in the Juliana Anicia codex when the names of the plants are given in Greek, Latin and usually Arabic, and you are asking me to identify these awful pictures."

One of the earliest illustrated printed herbals was the *Herbarium of Apuleius Platonicus* (printed in Rome about 1481) with Latin text. (This is not the Apuleius who wrote *The Golden Ass*). Virtually nothing is known of him, but his work is believed to have been originally written in Greek about 400 A.D. There are many extant manuscripts of it, including a finely illustrated translation into Anglo-Saxon about 1050 A.D., now in the British Museum.

But the three main herbal Incunabula, all originally printed in Mainz, are the Latin *Herbarius*—1484, the German *Herbarius*—1485, and *Ortus Sanitatis*—1491. This latter also appeared in translation as *Jardin De Sante* and *Gart Der Gesundheit*. There is a good facsimile of the German *Herbarius*. The woodcuts from all these are all rather stylized, obviously not drawn from nature.

Plate 27 is a typical illustration from the Latin *Herbarius*.

Plate 28.—I couldn't resist introducing this as a lighter note. It shows the woodcut of Narcissus from *Ortus Sanitatis* beside Lear's "Manypeeplia Upsidownia" from his *Nonsense Botany*.

Plate 29 was prepared from a copy of *Ortus Sanitatis* at Dumbarton Oaks, showing the female Mandrake picture and the text that goes with it.

A new era in the history of the herbal may be said to date from 1530 when the first part of *Herbarium Vivae Eicones* was published in Strassburg. This is known as Brunfels' herbal, and in it the plants are represented as they are. Plate 30 shows a woodcut from this book.

Throughout the texts of all the early printed herbals (and presumably their forerunners in manuscript) runs the theme of the four elements and their natures—fire, hot and dry—air, hot and moist—earth, cold and dry—water, cold and moist. For instance, in the German *Herbarius* one finds on nearly every page a sentence such as "A finds this plant of the second degree of coldness and the third degree of dampness, but B considers it of the third degree of coldness."

In general, the illustrations in the early printed herbals are limited to two or three collections of stylized woodcuts copied over and over again in more and more degenerate form. The same superstitions persist. Probably the most fantastic is the story of the Barnacle Tree. According to one version, trees sprout from the wood of shipwrecks,

shells grow at the ends of the branches, and from the shells small geese gradually emerge and fly away.

Plate 31 is taken from Gerard's herbal—1597. Gerard claims to have seen this process in action with his own eyes on a small island in Lancashire.

I make no apology for this short survey of herbals. I repeat that my purpose is to interest a wider circle, and the text and illustrations of herbals of the 13th, 14th, and 15th Centuries are a most interesting part of the background of this odd book. To the best of my knowledge no one has seen any book, certainly no illustrated book, of the period which covers the wide range suggested by the drawings in it.

There are many aspects of the study of the manuscript which I have not covered in this paper. Little has been said about Roger Bacon, very little of John Dee, nothing at all of such figures as Ramon Lull. There is as yet no solid evidence that the manuscript is not by Roger Bacon or a copy of a work by him.

BIBLIOGRAPHICAL NOTE

- [1] W. R. Newbold (edited by R. G. Kent), *The Cipher of Roger Bacon*, University of Pennsylvania, 1928. This book was published after Professor Newbold's death from his notes and would probably not have been published at all had he lived longer since, after the first announcement of his decipherment of parts of the Voynich Manuscript, Newbold made virtually no progress at all. In spite of the fact that his solution is now quite unacceptable, the earlier part of this book should be read by anyone interested as it provides a more comprehensive background than any other source.
- [2] Agnes Arber, *Herbals, Their Origin and Evolution*, 1470–1670, second edition, Cambridge, 1953. This is the standard work on the history of herbals.
- [3] Robert T. Gunther, *The Greek Herbal of Dioscorides*, Hafner, 1959. The text of this book is an early English translation of *De Materia Medica* of Dioscorides, very fully illustrated with line drawings traced from the illustrations of the Juliana Anicia manuscript.

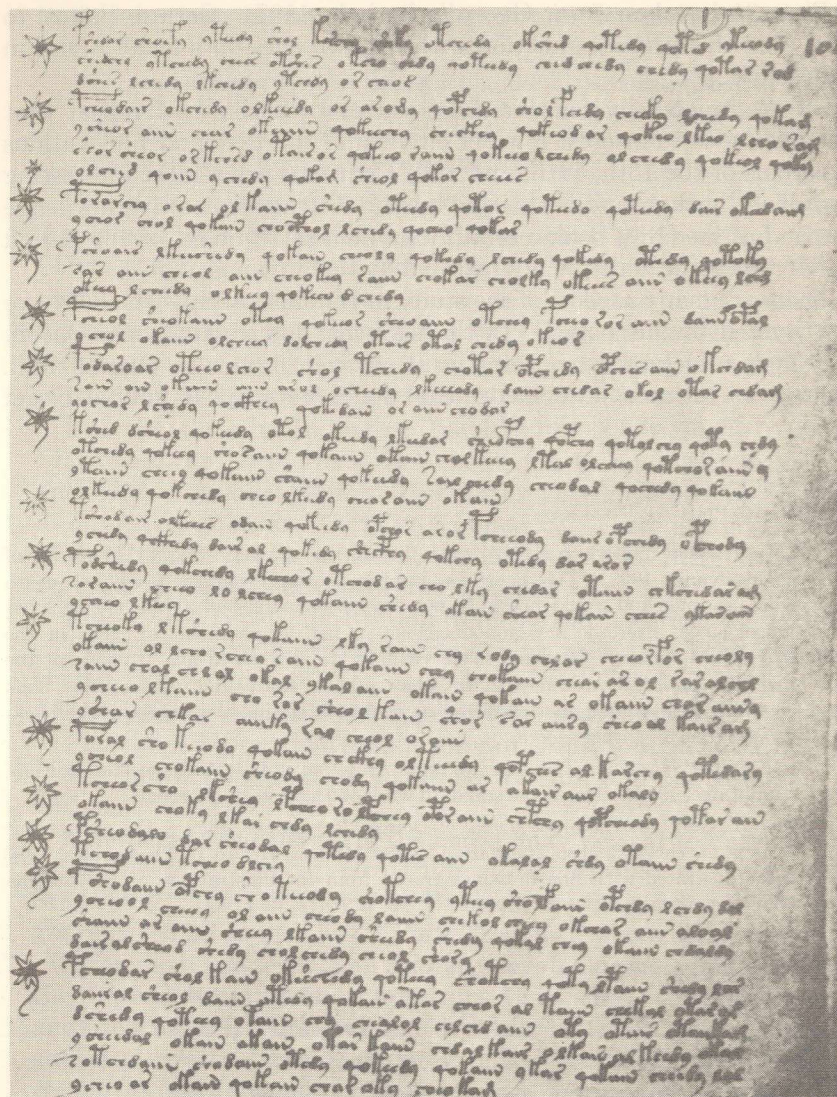


Plate 1.

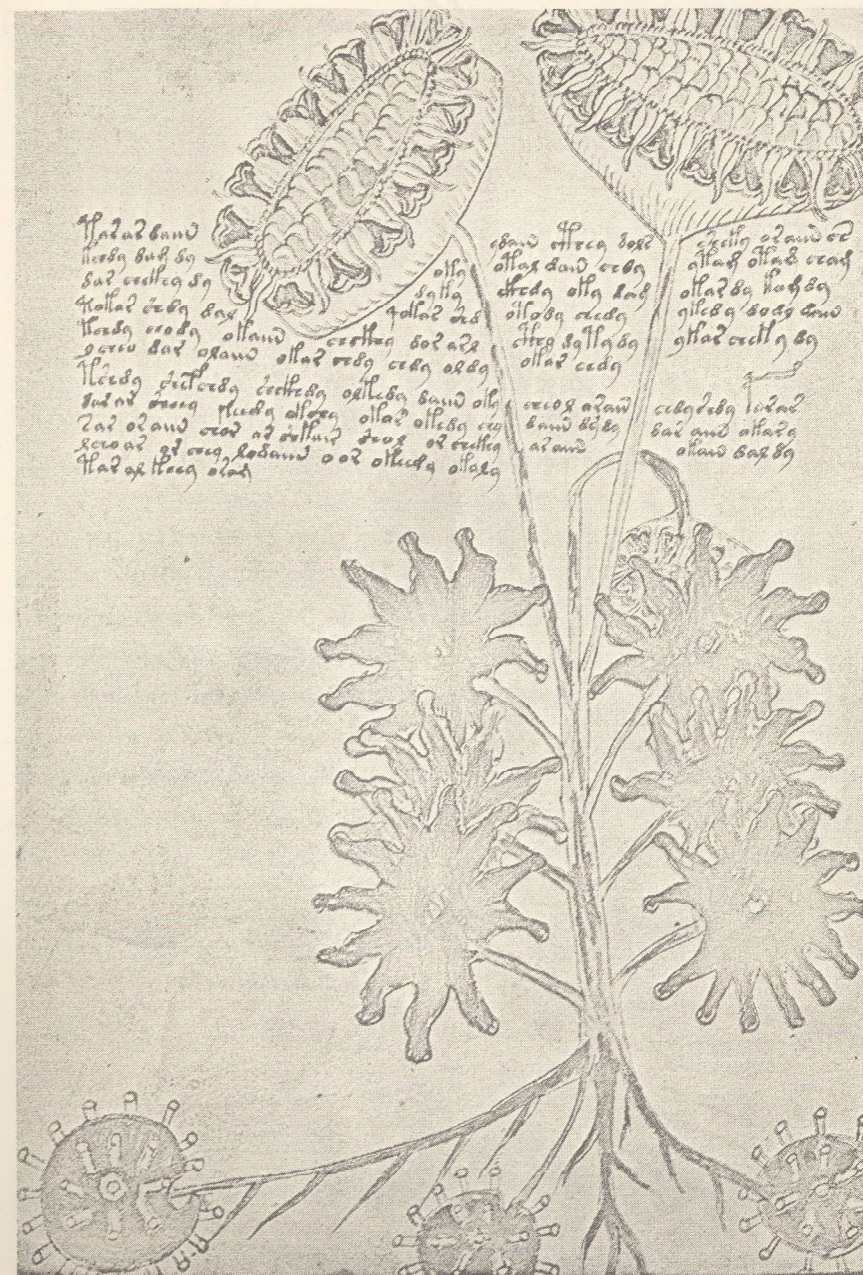


Plate 2.

Reuerende et Eximie Dñe
in Christo Pater

Librum hunc ab amico singulari mihi testamento relictum, mox
eundem tibi amicissime Athanasi ubi primum possidero cepi, animo des-
tinari: siquidem persuasum habui à nullo nisi attente legi posse. Petisti
aliquando per litteras ejusdem libri tum possessor judicium tuum parte
aliqua libri à se descripta et tibi transmissa, ex qua reliqua à te legi
posse persuasum habuit, verum librum ipsum transmittere tum reu-
cabat, in quo discifando posuit indefessum laborem, uti manifestum ex
conatibus ejusdem hinc una tibi transmissis, neq. prius huius spei quam
vires sua finem fecit. Verum labor hic frustraneus fuit. Siquidem non
nisi suo Kircheri obediunt quomodo ptinges. Accipe ergo modo quod
nunc tibi debebatur hoc quaecumq. mei erga te affectus indicium,
huiusq. lras si quas sunt, consueta tibi felicitate perumpo. redulit mihi
D. Doctor Raphael Terdinanti tertij Regis tum Diemina in lingua boemica
instructor dictum librum fuisse Ludolphi Imperatoris, pro quo ipse latoni
qui librum attulisset 600 ducatos protulerat, hethorem vero ipsum pu-
tabat esse Rogerium Baconem Anglem. ego judicium meum hic sus-
pendo. tu vero quid nobis hic sentiendum desine, cuius fauori et gra-
tis me totum commento manens.

Atq. vestre

Praga 19 Augusti.
4. 1008

Ad M. Marcum
Joannes Marcus Marci
in Cronland.

"THE MOST MYSTERIOUS MANUSCRIPT IN THE WORLD"

THE ROGER BACON CIPHER MANUSCRIPT

(BACON, ROGER ?.) Cipher manuscript on vellum. Text written in a secret script, apparently based on Roman minuscule characters, irregularly disposed on the pages. 102 leaves (of 116; lacks 14 leaves), including 7 double-folio folding leaves; 3 triple folio folding leaves; and one quadruple folio folding leaf. With added signature marks (of the XVth or XVIth century), and foliation (of the XVIth or XVIIth century) 1-11, 13-58, 65-73, 75-90, 93-96, 99-108, 111-116. With about 400 drawings of botanical subjects, including many of full-page size; 33 drawings of astrological or astronomical subjects, plus about 350 single star-figures; and 42 (biological?) drawings, most of which include human figures. The drawings colored in several shades of green, brown, light yellow, blue, and dark red. Large 8vo (c. 230 × c. 160 mm.). Old limp vellum covers (now detached). From the libraries of John Dee (?), the Emperor Rudolph II (reigned 1576-1611); Jacobus Horcicky (Sinapius) de Tepenecz; Joannes Marcus Marci of Cronland (1666); Athanasius Kircher, S. J.; and Wilfrid M. Voynich. Accompanied by an Autograph Letter signed by Joannes Marcus, presenting the book to Athanasius Kircher.

No place or date, (XVth century, or earlier?).

An enigmatic mediaeval manuscript, which for over forty years has baffled the scholars and cryptographers who have attempted to wrest its secrets from it. It has been termed by Professor John M. Manly, who made a detailed study of it, "the most mysterious manuscript in the world."

Plate 4.

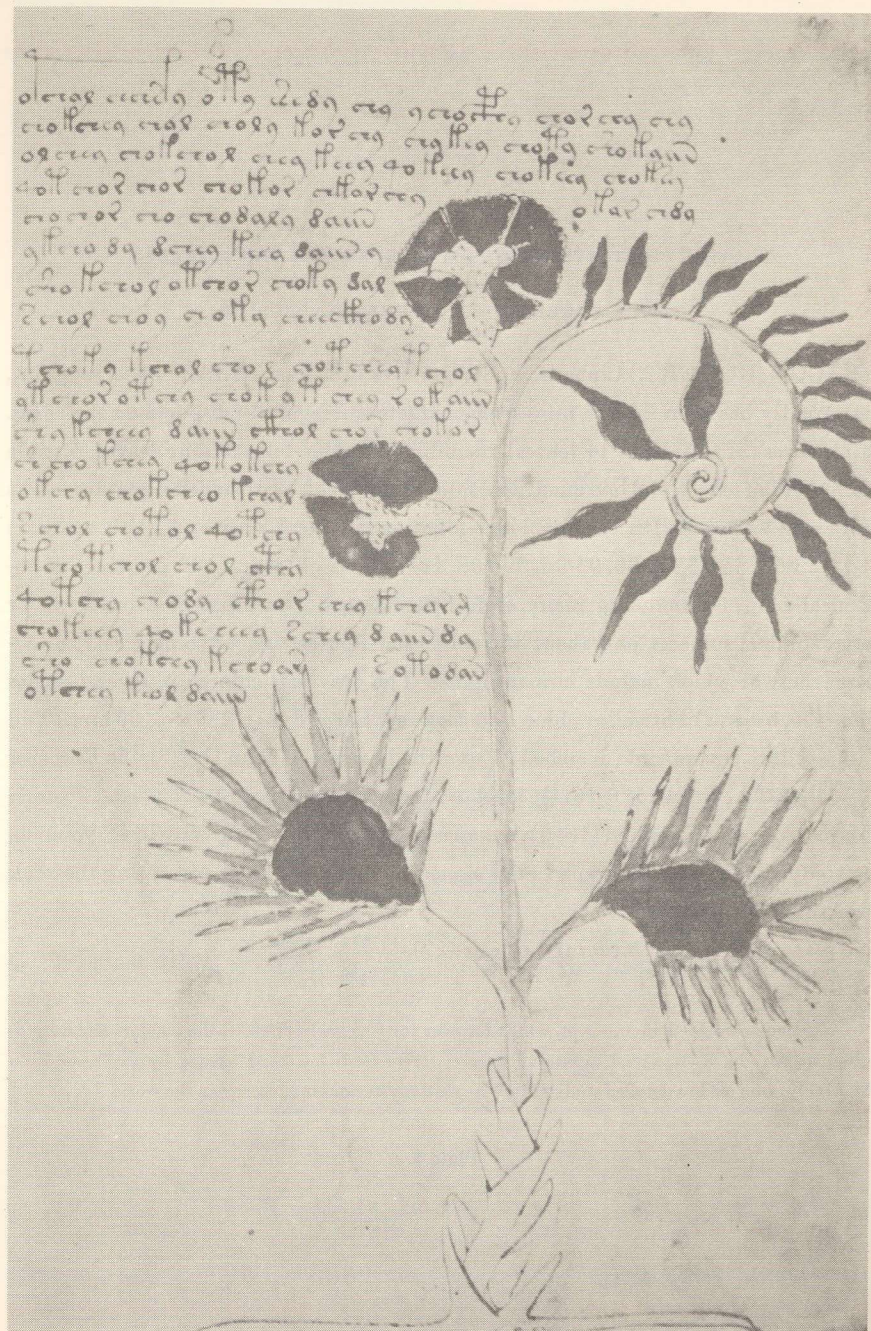


Plate 5.

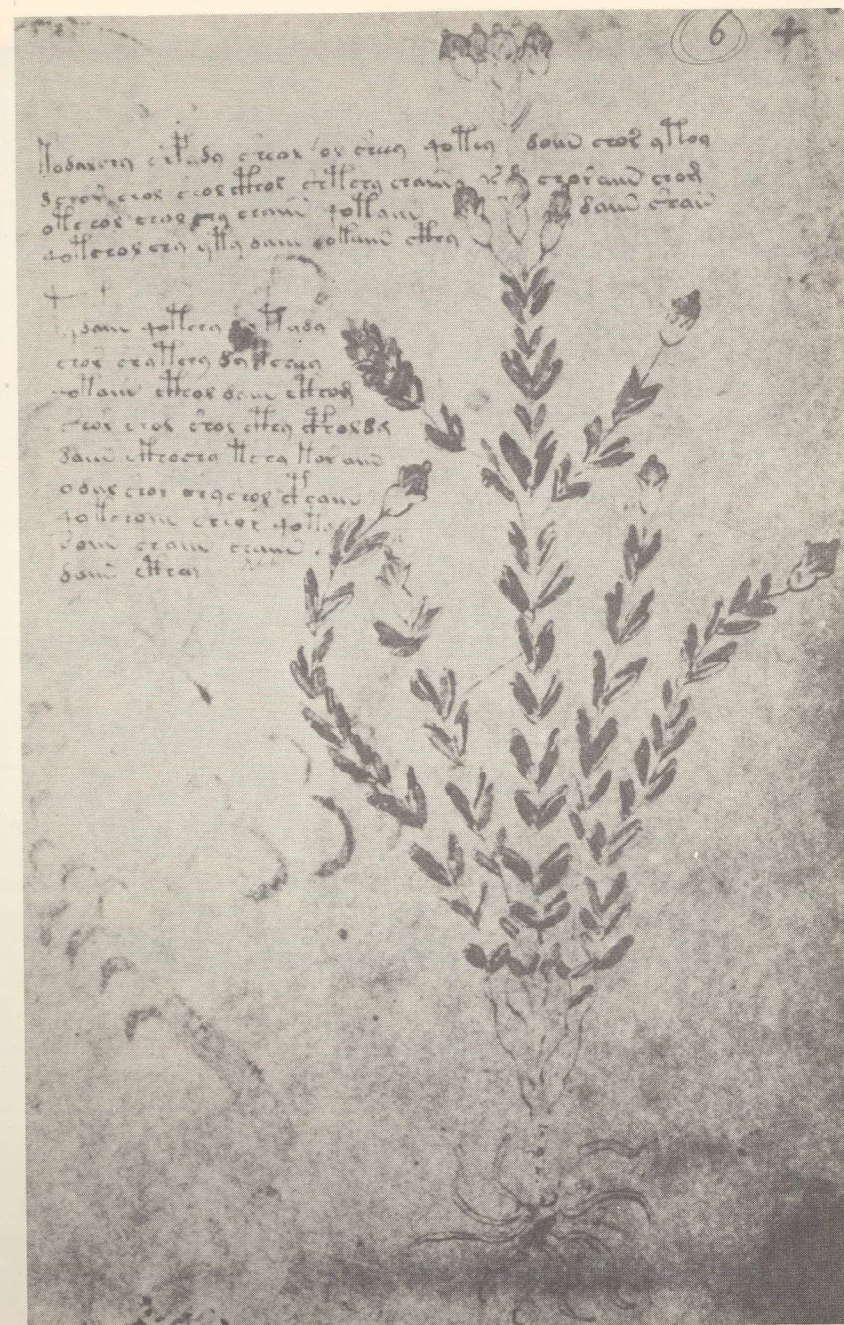


Plate 6.

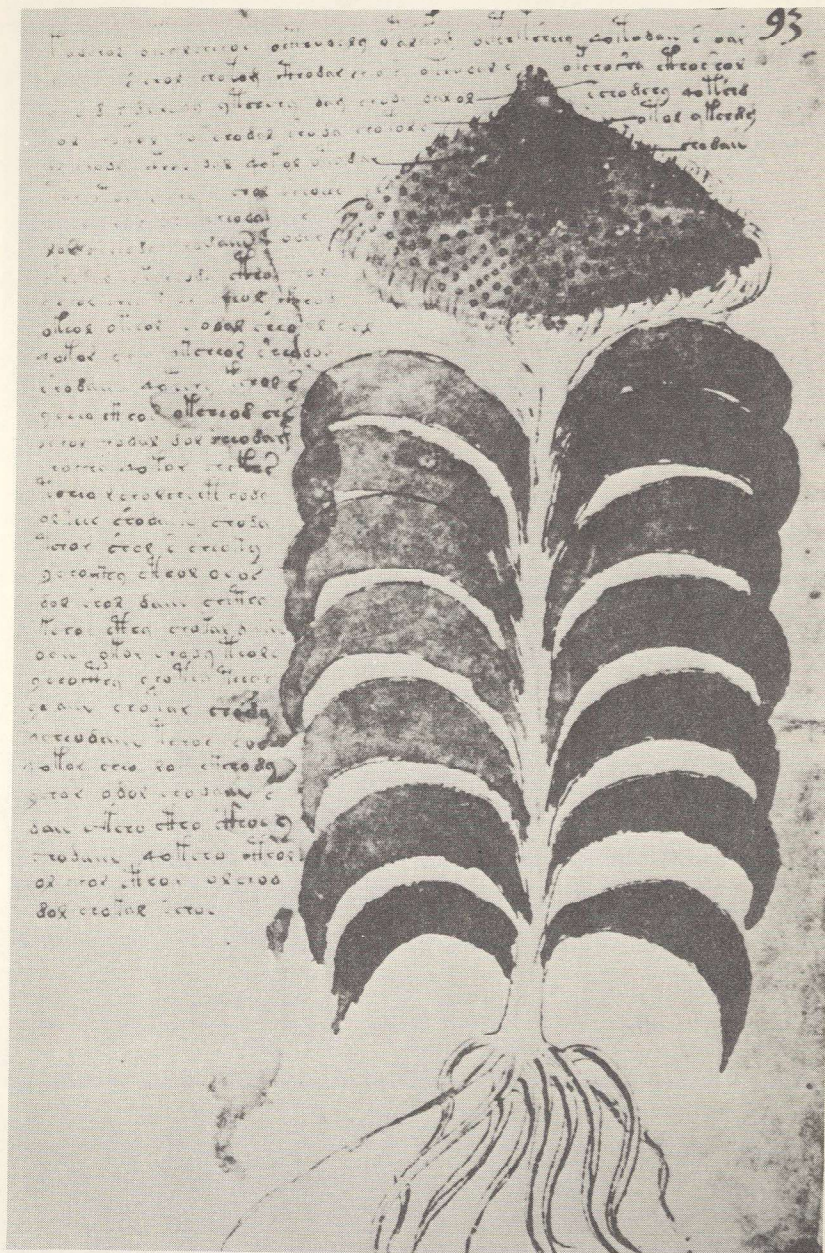


Plate 7.

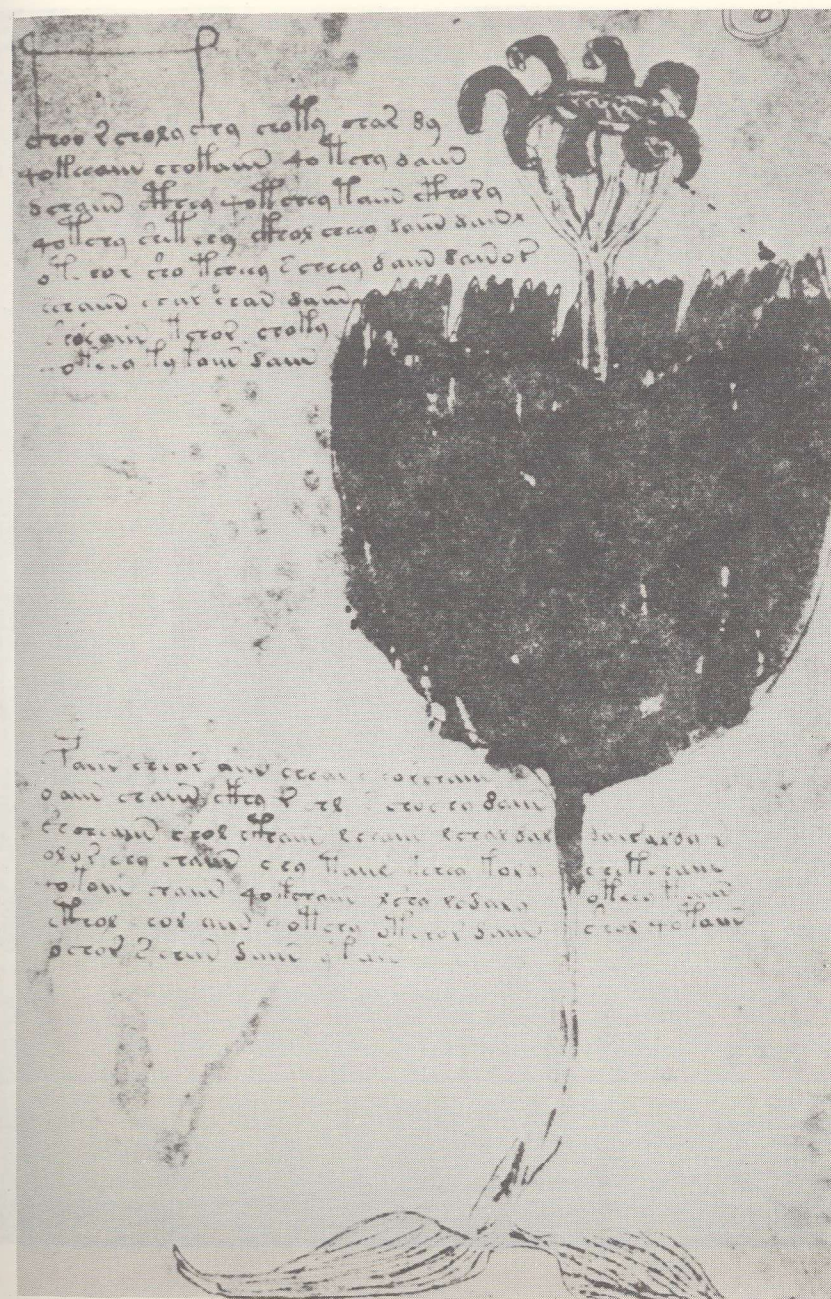


Plate 8.



Plate 9.

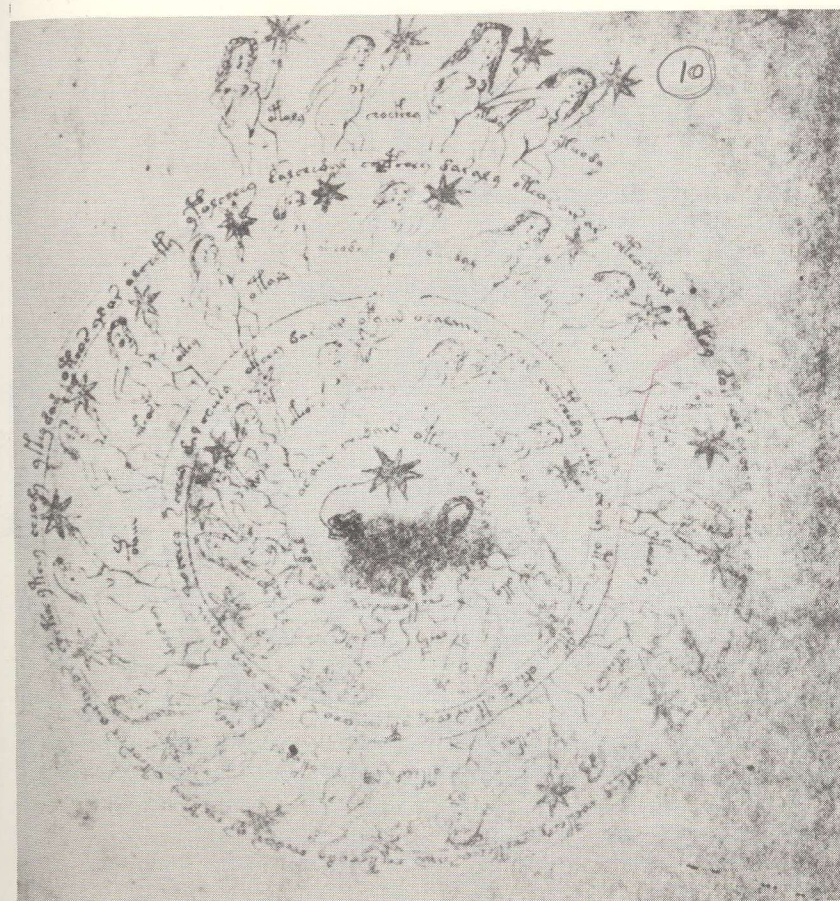


Plate 10.



Plate 11a.

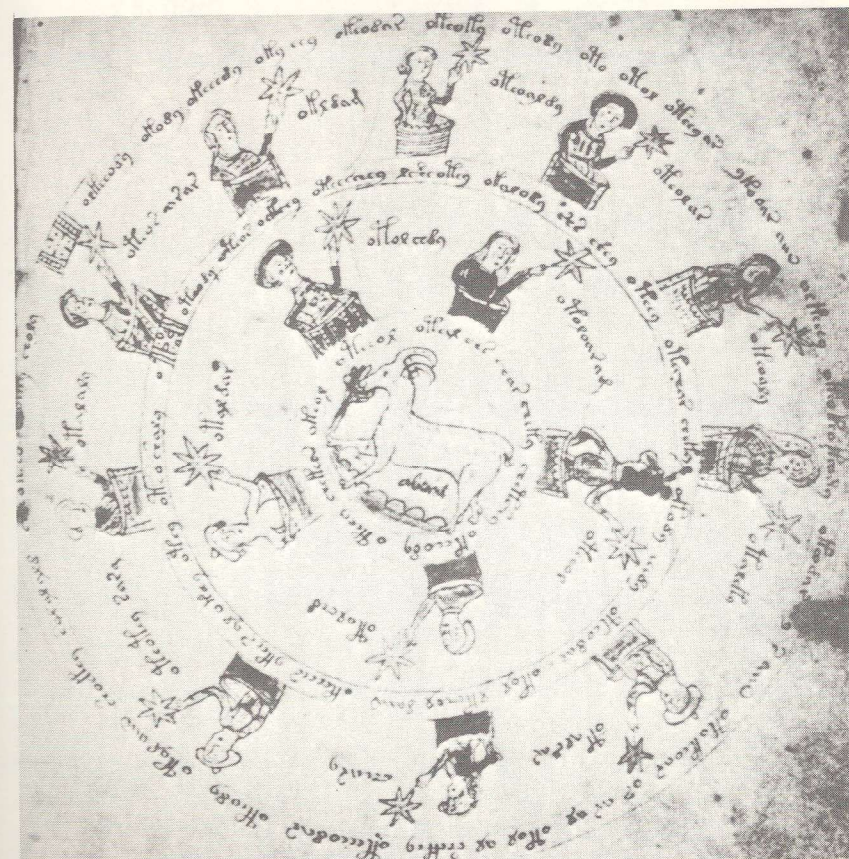


Plate 11b.

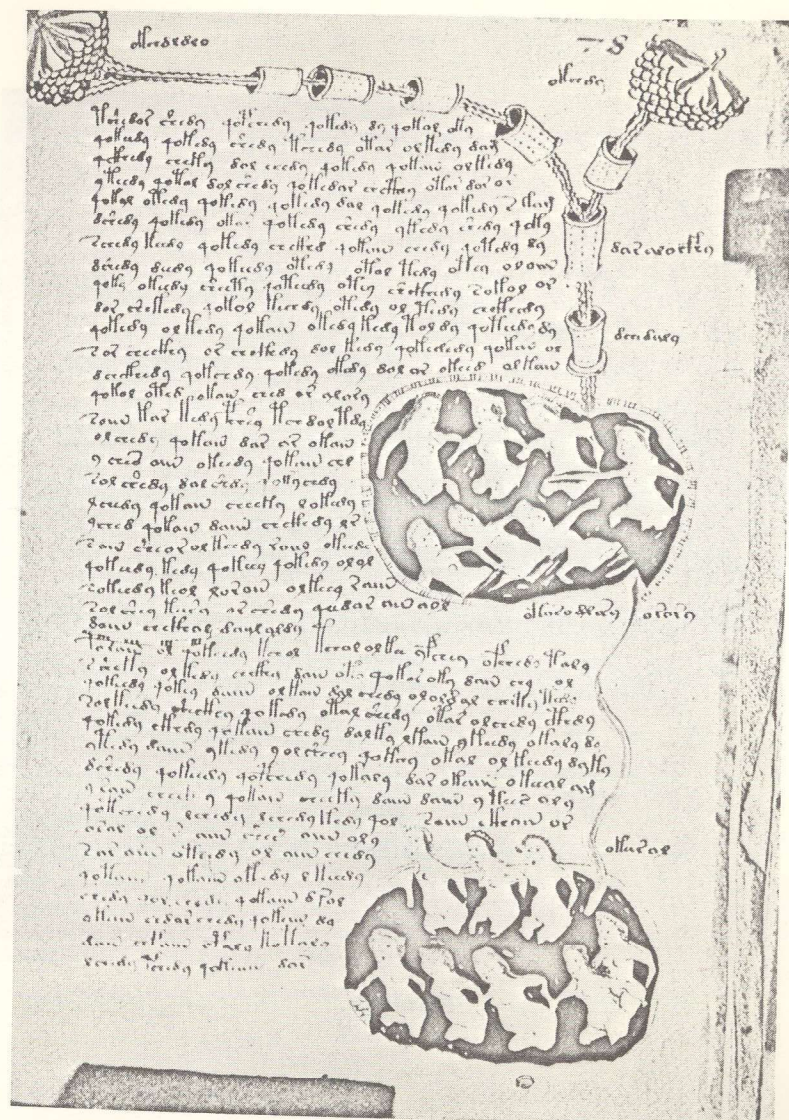


Plate 12.

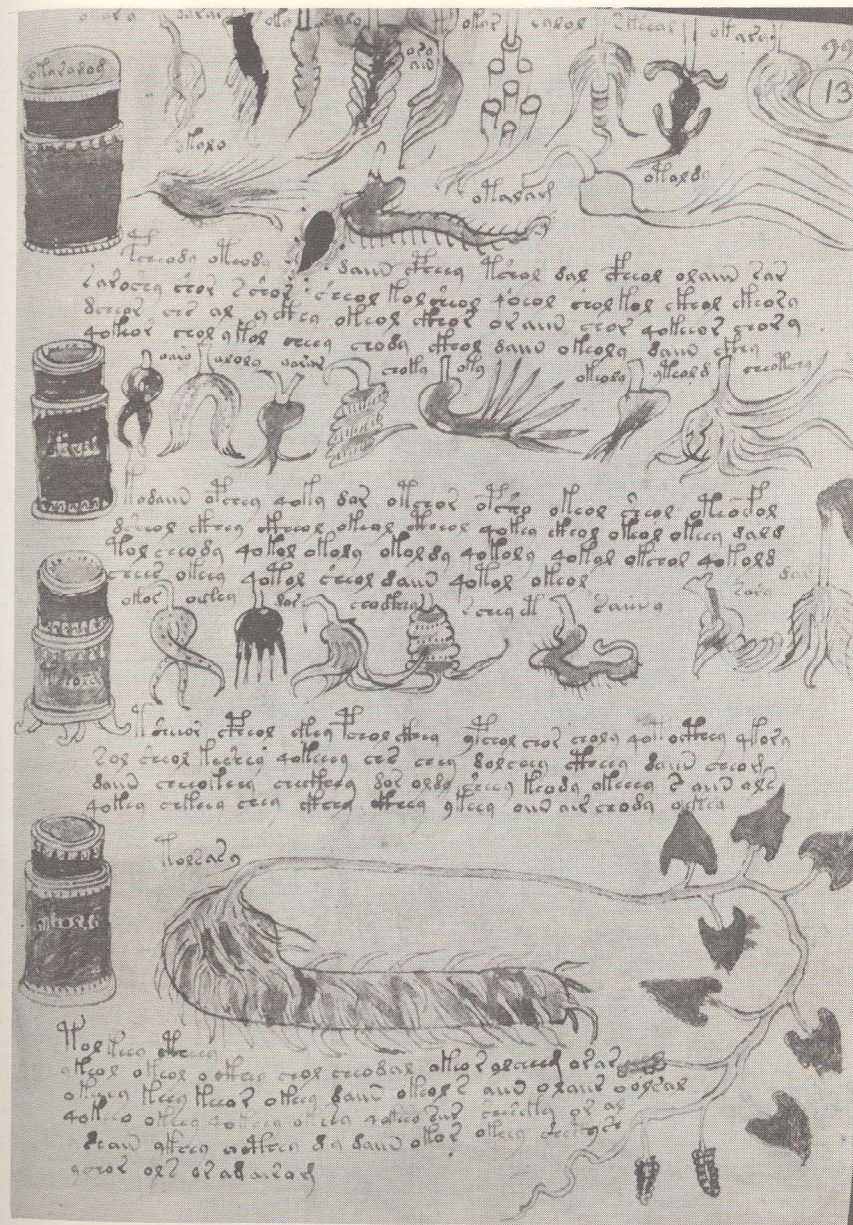


Plate 13.

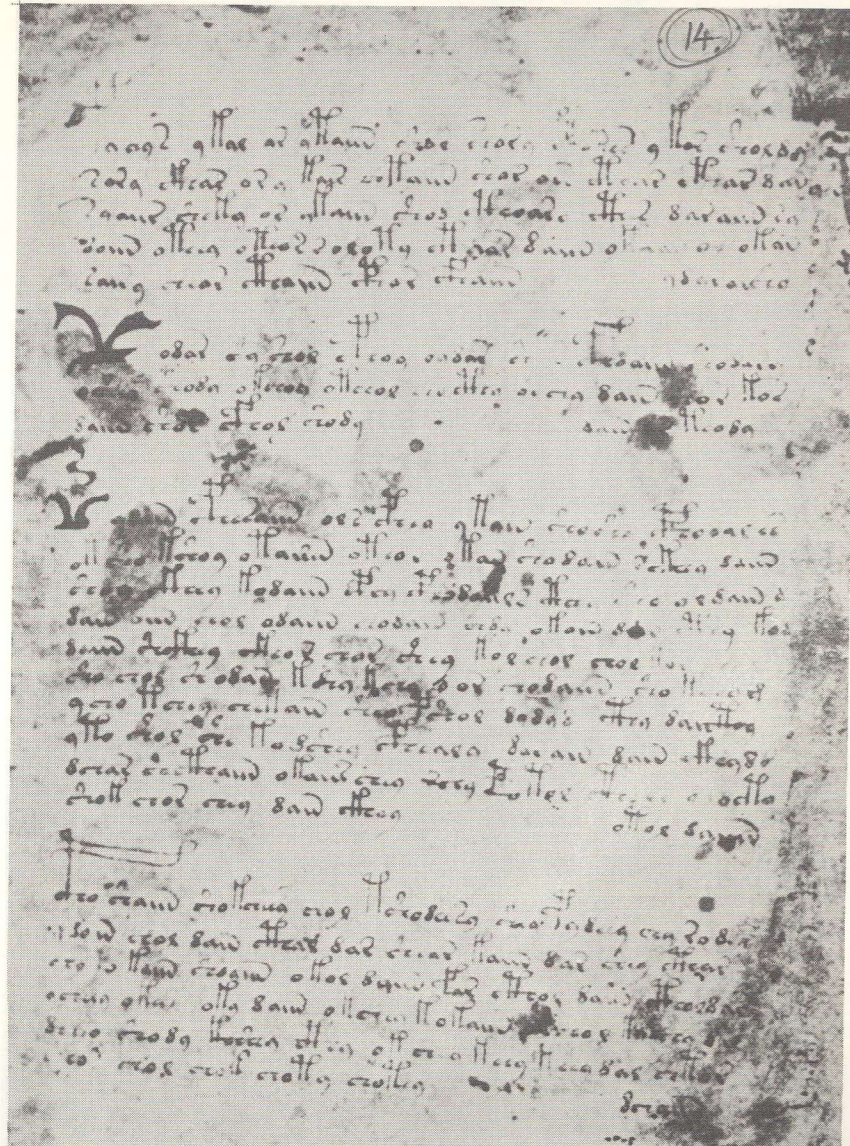


Plate 14.

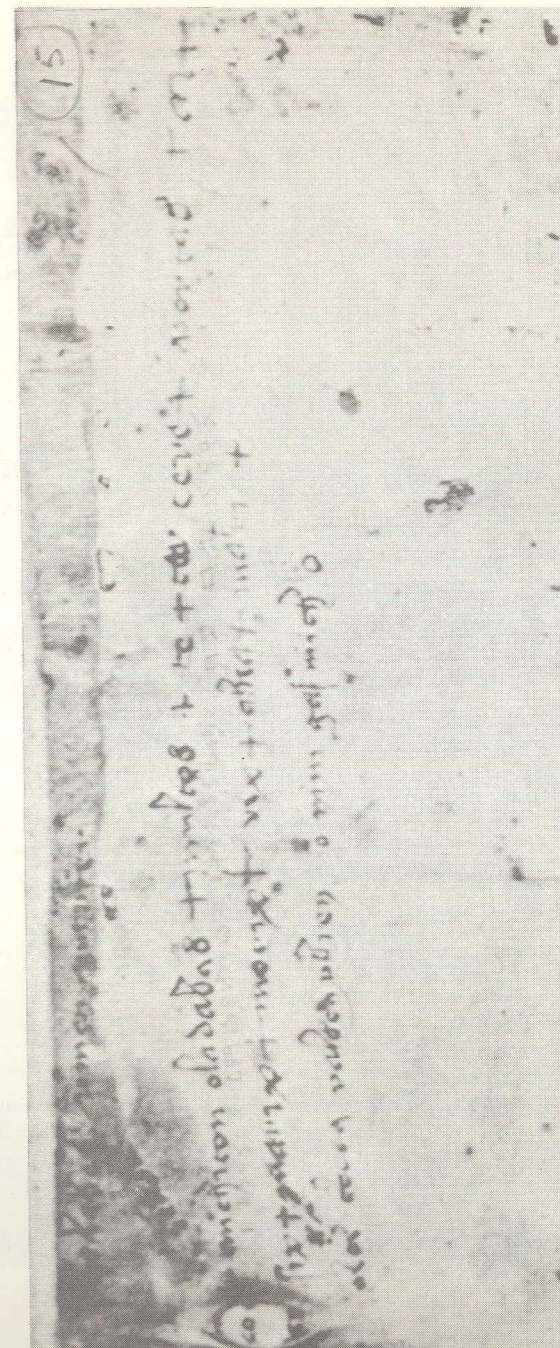
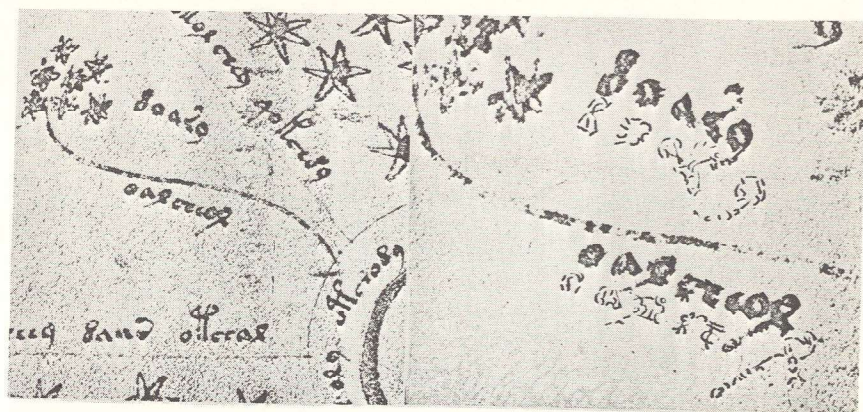


Plate 15.



Part of folio 68 recto, in various enlargements: A is enlarged about two diameters, and B about four diameters. After a certain amount of enlargement, the characters lose definition, but the drawings of the microscopic characters by Newbold in B look probable.

Courtesy of Wilfrid M. Voynich

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
M	M	9	8	2	4	0	a	c	\	cc	2	2	2	2	ffc	ffc	common symbols
~~M~~	~~M~~														~~ffc~~	~~ffc~~	
D	H	G	8	2	4	0	A	C	I	T	S	L	R	E	DZ	HZ	conventional

[illegible]

- (i) DOEA I I L S D C G 4 0 D A E E D C 8 G 4 0 D C 8 G
4 0 D C 8 4 0 H C 8 G O D A E T 8 A R . A E T A R . A I I L
- (ii) G T C C O R . A I I L O D C C G . T C O E T C 8 G O H T 2 8
O H T C 8 G A E D C 8 G A E . E D C A E . O E . 2 A I L . A E
- (iii) 8 A I R O D A E T C 8 G 4 0 D C C 8 G O H C O R A E
4 0 D C 8 G 4 0 D C 8 G O D A I I L O H C 8 A R .
T A E D C C 8 G
- (iv) G T C 8 A I L . O R . T C O 8 G 4 0 A I I L O D C C G
4 0 D C C G T 8 A E O D C 8 G 4 0 D C 8 G O D C 8 A E
T 8 G
- (v) 2 A I L A R . O E D C C G 4 0 A E . O E D C 8 G O D A I I L
- (vi) H O E T C 8 G O D C 8 G 4 0 H T C 8 G O D C O E D C G
H C C 8 G 4 0 H A I I L O H C 8 G O D C C 8 G
O H C 8 A E
- (vii) O D C O E T C 8 A E T O D C 8 G O 4 0 D C 8 G O H T C 8 G
2 A I I L . A E D A L

Preliminary division of very common words into "roots" and "suffixes" :-

Roots:- OD, OH, 4OD, 4OH, T, S, 8, 2, ED

Suffixes:- either (i) c, cc, ccc followed by G or 8G

or (ii) one or more of the following :-

AL	AIL	AILL	AILLL
AR	AIR	AIRR	AIRRR
AE	AIE	AIEE	AIEEE
OR	OE		

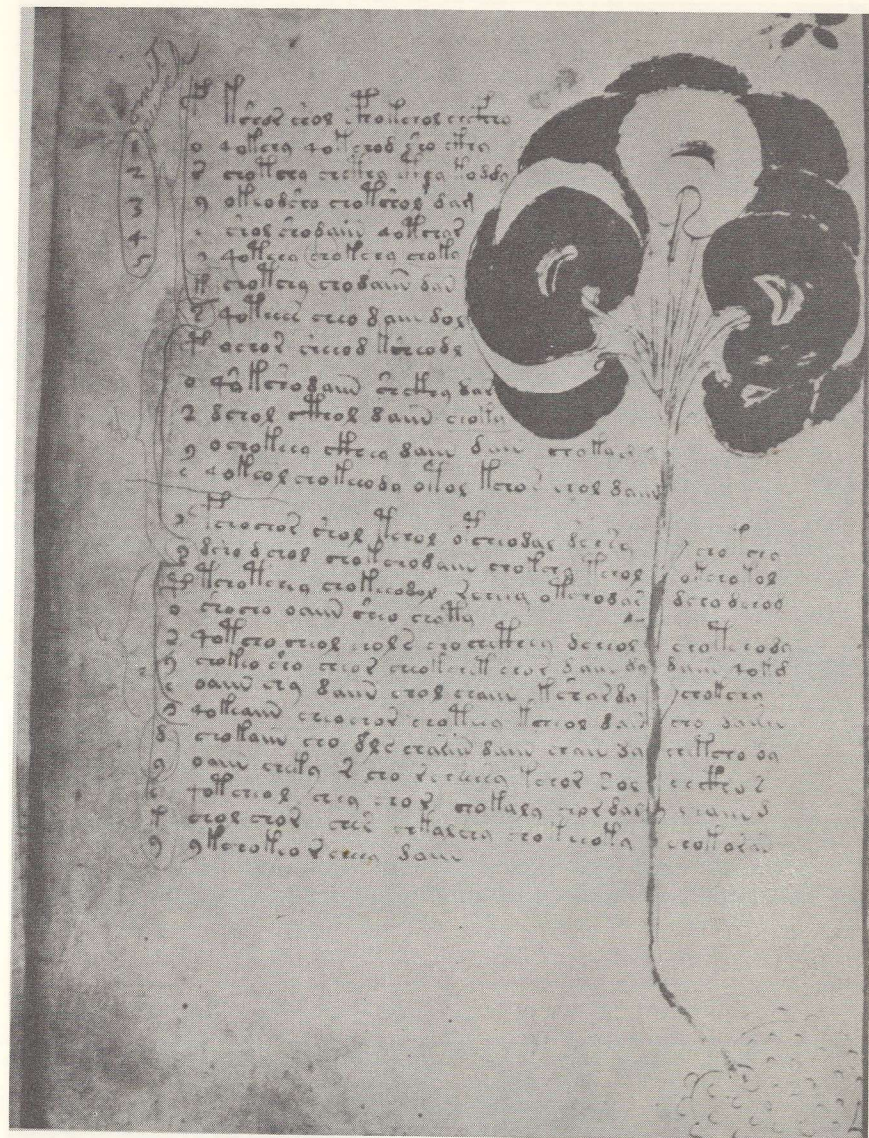


Plate 18.

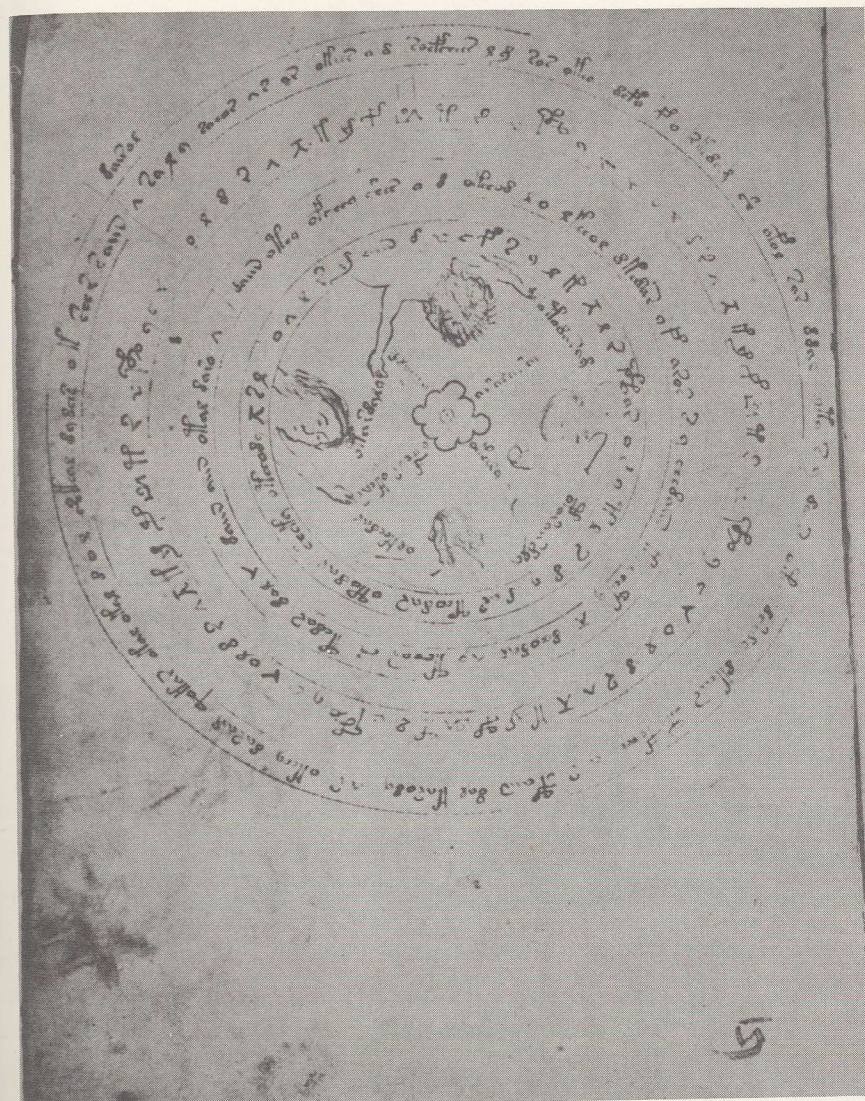


Plate 19.

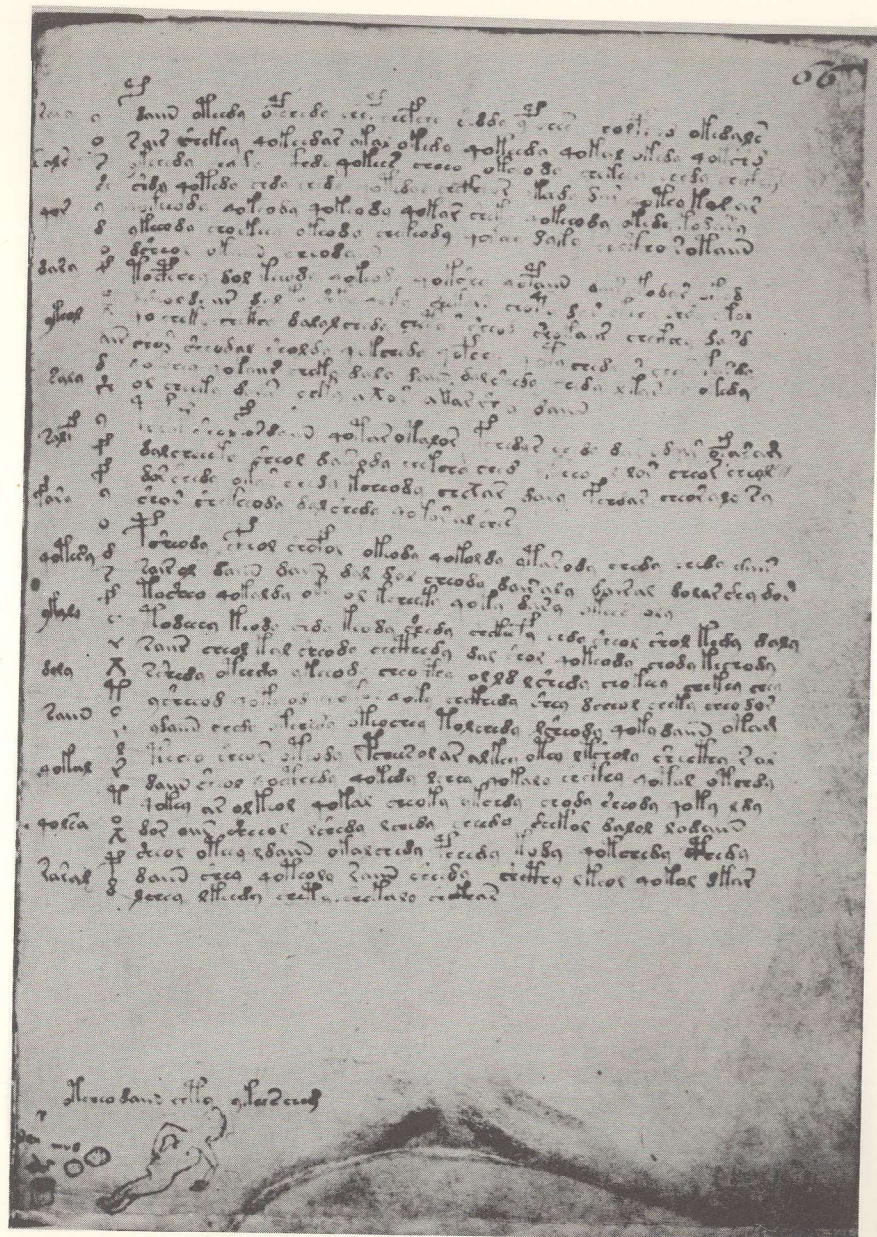


Plate 20.

THE UNIVERSAL CHARACTER

By which all the Nations in
the World may understand
one anothers Conceptions,
Reading out of one Common
Writing their own
Mother Tongues.

A N. I N V E N T I O N Of GENERAL Use,

The Practise whereof may be Attai-
ned in two Hours space, Observing the
GRAMMATICAL Directions.

Which CHARACTER is so contri-
ved, that it may be Spoken as well
as Written.

By Cave Beck, M.A.

L O N D O N,

Printed by Tho. Moxey, for William Weekley, and
are to be sold at his Shop in Ipswich. 1657.

Plate 21.

that conjunct	ina	to threaten	r 1215
to thatch	3757	to thresh	r 1923
to than, no	1261	to thrive, v. grow	r 2016
a theater, tam	480	thrift or frugality	r 2016
a theefe	p 1745	the throat	r 874
then adverb of time	fni	a throne	r 2430
then in comparison	fno	to throb	r 116
thence	fnu	to throng, v. Crowd	r 1399
Theology	r 1606	a threstle bird	r 3769
therefore	ipa	to throw or cast	r 927
there	fne	thrummies	r 3770
a Thieft	3182	a through bird	r 3771
thicke	q 802	to thrust	r 1400
the thigh	p 3758	a thumbe, +1,	r 1907
a thiller of a Cart	r 3759	to thumpe or knocke	r 670
a thimble	r 3760	a thunderbolt, furc	r 624
thinne	q 312	thunder	r 624
thine	he	Thursday	+5, r 1484
a thing	r 2069	to thwack, ant	r 508
to thinke	r 1163	thyme herb	r 3772
to thirst	r 1654	to tickle	r 3773
this	ho	the tide	r 3774
a thistle	r 3761	tidings	r 3775
thither	spe	to tye or binde	r 302
a thong	r 3762	tiffany	r 3776
a thorne	r 3763	a t-ger	r 3777
a white thorne	r 3764	a tick of a bed	r 3778
a blacke thorne	r 3765	a tick worme	r 3779
a thorne-backe fish	r 3766	a tile	r 3780
a thorpe or village	r 3767	a gutter tile or rooffe tile	r 2197
through	per	to till or allure	r 138
thou	e	to till the ground	r 1422
thraldome	r 626	till or untill	spu
thread	r 3768	a till	
threadbare or old	208		

Plate 22.

An Example of writing and speaking the fifth Commandement.

Honour thy Father and thy Mother

Write { leb 2314 p 2477 and pf 2477

Speak { leb toreónfo, pee tofósénfen, & písf tofósénfen

That thy days may be long

fna, her 14848 mb 1716

fna heronfórafos, mib onsenónsic

In the Land which the Lord

in p 1699 vip 2529

in ponticinnin vip tohtónin

Thy God giveth thee

hep 306 pce b 510

hep tréofic pee befóno

* Note, for Euphony sake or the better sound, the letter [t] in [tre] may sometimes be left out and [t] in [at] may for the same cause be omitted or changed into the Consonant following as onfórafo or onforaffo, which I berry is usually taken in all languages as Commend for Commend assemble for assemble.

L A U S D E O.

Plate 23.



Plate 24.

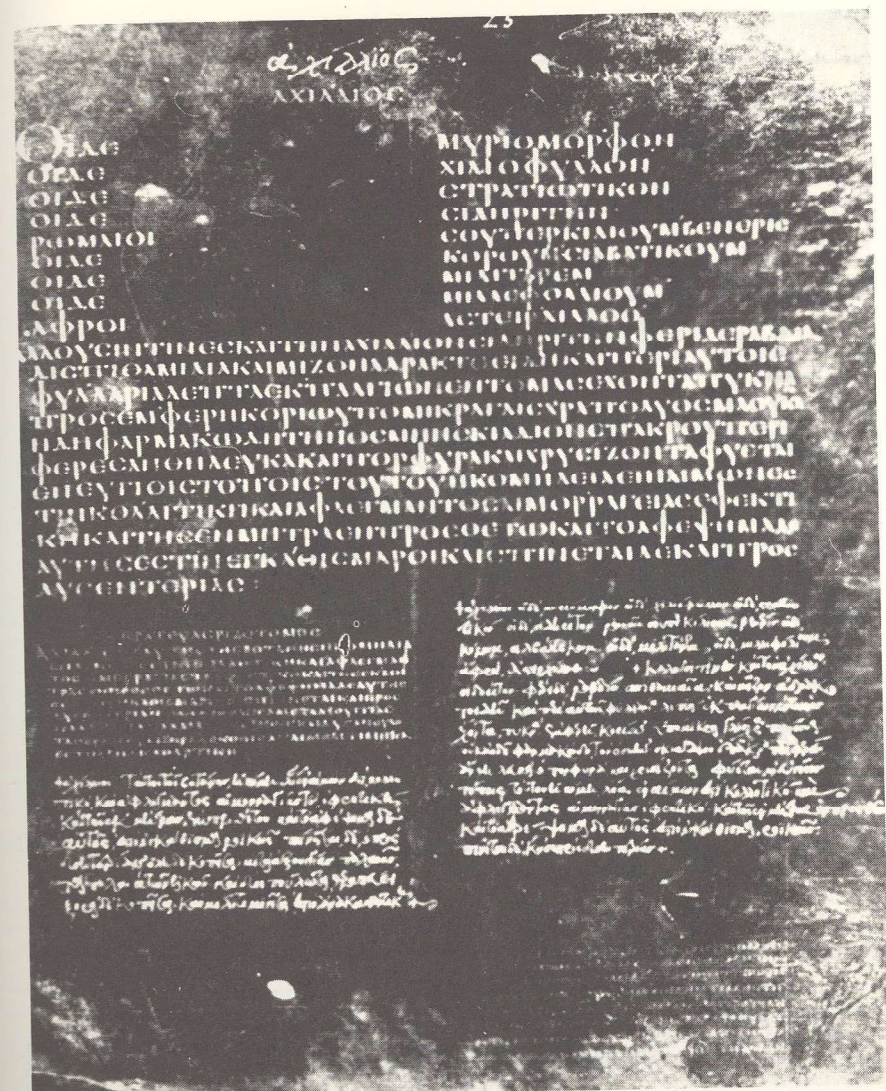


Plate 25.



FIG. 3.—TRACING FROM THE JULIANA ANICIA CODEX OF 512, FO. 5 v.

Epinoia holds a mandrake in her hand. Krateuas is painting it, while Dioskurides writes an account of it in a codex. In the original figure much of the paint has scaled off and our restoration is conjectural in places.

Plate 26.



BRIONIA

Fig. 6. "Brionia" [*Latin Herbarius* (Arnaldus de Villa Nova, *Tractatus de virtutibus herbarum*), 1499]

Plate 27.

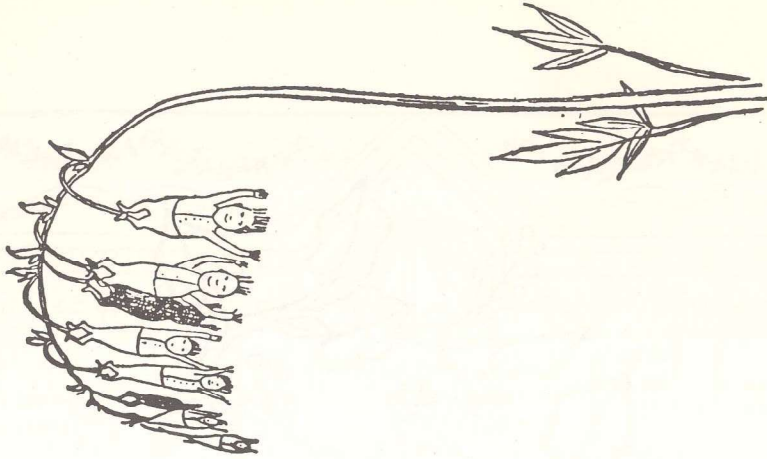
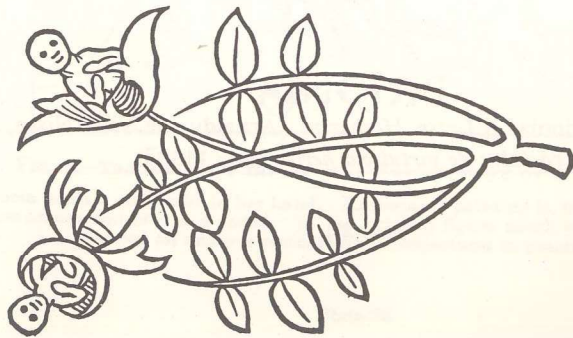


Plate 28.



Tractatus

Alia species q̄ nōlatur narbus ad
ministrat a chirurgicis q̄n volunt membri
aliquid incidere. et q̄n bibit solat q̄d d̄r
suffocās ei est tyriaca. Et idē auc.
Rasis. Dixit mihi q̄dā ex antiq̄s babilo
nie. q̄ q̄dam puella p̄medit quinq; poma
mādragore. et cecidit sine cop̄iata. et tota
efficēta est rubicūda. et quidē su p̄uicēns
effudit su p̄caput ei⁹ aq̄ minis donec sur
rexit. P̄ Ego vidi hōi ei⁹ sumple
runt de radice eius cā impingūā di. et ac
cidit eis sicut accidit solz hōibus ingredi
en tibus balneum et bibētib⁹ post exitum
vinū multū. nā facit suire vulnis eoz nimis
rubicundus. P̄ Et idē auc. dyal.
Radice mādragore multi dāt ad amorē.

folior eius est similemēspili et est losach.
et i ciniri colo. h̄ns odorē bonū. et terra ip̄s
sunt grana similia granis p̄troz. et habet
radices magnas mediocriter duas l̄tres
adherentes inalcē q̄ren⁹ nigras et inter⁹
albas. sup̄ q̄s est cortex grossus. Et hec
species mandragore non habet stipitem

Operationes.

Et Mandragora fortissimi odoris est.
ab hōi incicuno nō colligit. P̄ Utri
usq; vis una est. Hec cum poltra n̄tra fer
uores oculor et dolores aurū sedat. C
Radix eius cū aceto trita et illita ignem sa
crum curat. P̄ Tricenna. Mandra
gora somnū p̄uocat. Et q̄n ponit in vino
vehementer inebriat. Adulm̄q; visus ei⁹
et odoramentū. facit apoplexiā. Et
Lac eius enclit lentigines. et pannū sine
mordicardē. Solvendo it̄ eductio d̄r
legma. P̄ Radix eius trita et m̄lita
ro imposita sup̄ benisipilam sanat eā. Se
men eius mariscan mandificat. et vomit
um p̄uocat.



La. cclxxvij.

Mandragora femine. Scraḡ. auc.
dyal. Et semine color est niger et no
minus aurandachis sine bādachis
m̄lita m̄ca. P̄ā in folijs ei⁹ est similitudo
cū folijs lacine. et sunt pinguis q̄ntis odo
ris. et exundant sup̄ faciem terrē. In medio

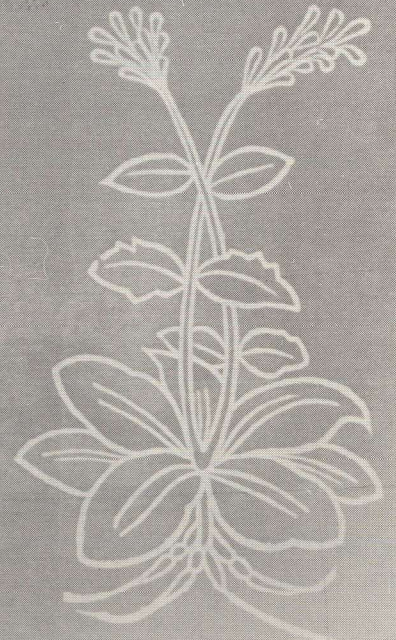


Plate 29.



Fig. 22. "Walwurtz männlin", *Symphytum officinale* L., var. *purpureum* Pers., Comfrey [Brunfels, *Herbarum vivae eicones*, vol. 1, 1530] Reduced

Plate 30.

The breede of Barnakles.

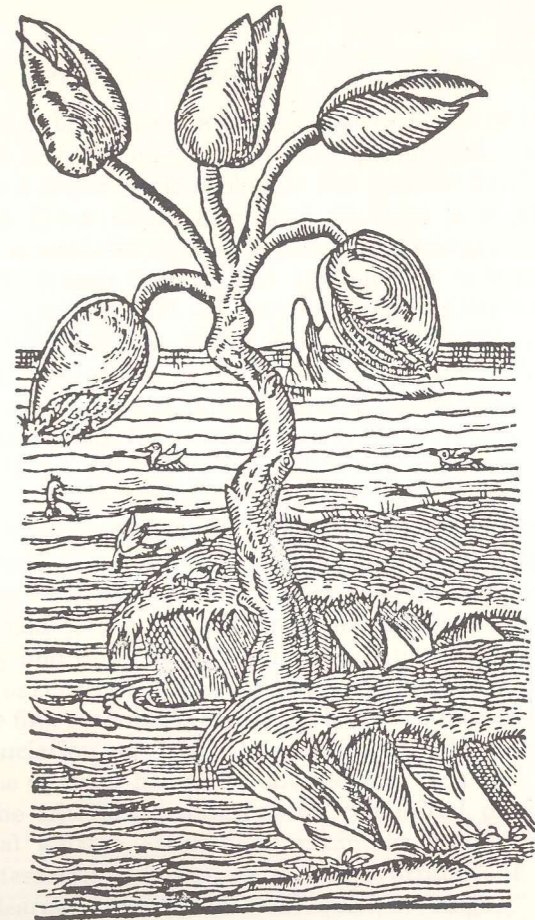


Fig. 60. "The breede of Barnakles" [Gerard, *The Herball*, 1597]

Plate 31.

Linguistic Canonical Forms: Tools for Analytic Progress

BY CLIFTON H. BUTCHER

THE LAY CONCEPT OF LANGUAGE

The lay idea of language in a literate society is one of "words," their meanings, and their spelling. The word is paramount—and publishing houses turn out paper back books for the popular market with titles such as *Word Power Made Easy* and *30 Days to a More Powerful Vocabulary*. Success for a typist is partially predicated on typing correctly, i.e., the correct spelling and hyphenation of words as well as their correct punctuation in sentences; therefore there are books such as *Spelling for the Millions*, *Words Most Often Misspelled and Mispronounced*, and *6 Minutes a Day to Perfect Spelling*. And there are similar guides pertaining to style and word selection for writers and editors (*English Verbal Idioms*, etc.).

Word selection is the key to success; choose the exact word with the precise "meaning" to portray the "thought" you are trying to convey! And in the American society there must be an "authority" (an Emily Post as it were) for those meanings—and the authority is the dictionary. Any dictionary at hand is used to settle any argument. No attention is paid to the initial date of compilation, the method used for selectivity and ordering of meanings, or to whether the meanings represent actual current usage rather than egocentric or dialectocentric beliefs. The first meaning is taken as the prime meaning and the first listed pronunciation as "the" pronunciation! If a word has five meanings, one of them has to be entered first in a line of print—but what were the criteria for selection for the initial position? Was it first historical written attestation; meanings in alphabetical order; frequency attestation in literature; meanings arranged in an order from formal and learned "down" to "colloquial," "vulgar," and "slang"; meanings arranged in regional order with New England usages coming first; or random order? There is a dictionary for every price range and every taste—and the same holds true in selecting a wardrobe, a car or a house.

For most "amateurs of language," i.e., "lovers of language," who often get degrees in language and literature and then, to earn a livelihood, become teachers of said subject matter, or turn to the business world or government and become travel agents, translators, researchers,

foreign language copy writers, interpreters, or possibly "analysts" using language as a tool, the above lay concepts are still paramount. They see language as being "thoughts" expressed in "complete sentences" which are spelled and punctuated "clearly"—with the key to correct expression being word selection according to "basic" meanings of words (with basic having something to do with etymology or formal writing and nothing to do with frequency of usage by the speakers of the language).

With this underlying concept of language it is easy to see why the "analyst" believes refinement of his language tool consists in the preparation of more and better specialized glossaries of words and their meanings, and in frequency counts of words, letters, digraphs, and trigraphs. Such counts are made by "dragging" one or more dictionaries or are based on running text (literature, newspapers, scientific periodicals, textbooks, *inter alia*) All such counts vary appreciably. Counts based on dictionaries are the worst since they reflect each and every lexical entry (in non-inflected form) whether obsolete or current and bear no statistical relationship to content or style of any given text. Counts based on running text are biased with the degree of bias reflecting the topics and styles of the selected texts; such counts are invalid to varying degrees when applied to other types of traffic or subject matter.

THE SCIENTIFIC CONCEPT OF LANGUAGE

Language, according to most schools of linguistic science, is a structured interplay of a number of linguistic levels—with each level having its own basic linguistic building blocks which combine in various sets of relationships. A building block is not necessarily restricted to one and only one level. The most basic building blocks are the PHONEME, the SYLLABLE, and the MORPHEME. Phonemes are acoustic bundles perceived through or by a structural mental matrix as syntactically arranged in meaningful units called morphemes. Syllables are the carrier wave of human speech; phonemic distinctions associated with the "meaningful" elements of speech are mapped on to the carrier wave units by articulatory modification of the vocal tract. Morphemes are phoneme sequences which bear a "meaning" (morphological, lexical, syntactic, psychological, social, and/or "semantic") pertinent to the linguistic conventions of some given language and society; these morphemes enter into syntactic arrangements known as roots, stems, affixes, "words," idioms, phrases, sentences, paragraphs, topics, discourses, and even works of art (literature) and commerce (newspaper, books).

Some languages have "words" in that discrete phonological or syntactic markers are habitually associated with a certain low level of

lexicality. English is one of the many languages in which "words" are *not* a structural unit—our so-called words are psychological correlates to several types of linguistic structural units; as psychological units these words are the legitimate field of study for a number of disciplines that are concerned with various aspects of semiotics (anthropology, sociology, psychology, psycholinguistics, sociolinguistics, philosophy, religion, politics, advertising, etc.). Investigators in some current theories of linguistics are investigating formally, rigidly and scientifically certain aspects, properties and fields of semantics, establishing testable hypotheses to account for data and subsets of data which bear on the interrelationships between the sound units of physics, the patterned activity of speech habits, and the extremely complex multi-dimensional array of "meanings."

Put in another way, language, like practically all human behavior, can be described as particle, wave and field in a complex hierarchical ordering of levels. Acoustic elements of speech sounds are arranged into particles called phonemes—and these particles in linear sequence in time are sound waves (syllables and sequences of syllables)—and certain lengths of these waves are accompanied in a given language by conventional meaning correlates which can be analyzed as distributed in various sets and fields.

THE SYLLABLE

Put in still another way, for a given language a finite number of phonemes occur in a quite restricted number of phoneme sets each usually characterized by a vocalic nucleus, and these latter sets are known as syllables. A syllable is a time-linear sequence of one or more phonemes, and there are restrictions as to the total length of such a sequence, as to which phonemes can occur in which specific locational slots and as to the possibilities of adjacent co-occurrence of some or all of these phonemes. Syllables are the pulse-trains on which are superimposed structural and psychological elements such as affixes, roots, stems, "words," idioms and strings of same.

In writing, a secondary derivative of speech, the above structural and "meaningful" units are represented by sequences of letters (or other marks on paper or stone such as ideographs and syllable-graphs) which reflect on an imperfect non-one-to-one basis some, but not all, of the segmental and suprasegmental phonemes of a language. In ordinary lay terminology letters are divided into two classes, consonants and vowels. These letters are the minimal units of a partially rational stylized spelling system for the artifacts of culture called words. Letters are to linguistic science as the molecules are to physics in this age of atomic and sub-atomic physics: a convenience.

In linguistic terminology those phonemes and phoneme sequences which occur as syllabic nuclei are called vocalic or vowel phonemes, and those which occur at the onset and offglide margins of the syllable are consonant phonemes. For the linguist the minimal carrier wave communication unit is the syllable: a vocalic nucleus is obligatory and there are optional consonant phonemes or sequences at the linear margins of the syllable. In reality the phonemes are internalized in a mental matrix and are manifested in the world of physics by acoustic correlates actuated by articulatory movements modifying the shape of the vocal tract as various sound emitters or resonance chambers are switched on or off.

CANONICAL FORMS

Let V stand for any vocalic nucleus and C stand for any consonant phoneme or phoneme sequence. The generalized so-called canonical form of the syllable can now be symbolized as CVC; or, more truly, as (C)V(C) with the parenthesis setting off optional elements. Any structural unit can be represented by a symbolic canonical form and the linguists deal with a variety of such forms (as do mathematicians and logicians) at the higher abstract levels of linguistic theory as hypotheses are posited for testing and evaluation.

The symbol C could theoretically be expanded to an infinite number of C's, but the maximal consonant phoneme clustering at the onset and offglide margins of the syllable is sharply restricted in all human languages. The canonical form of the syllable in English is CVC; the expanded form is (C)(C)(C)V(C)(C)(C)(C). This expanded form accounts for 20 patterns:

V	CVC	CCVCC
CV	CCVC	CCVCCC
CCV	CCCVC	CCVCCCC
CCCV		
VC	CVCC	CCCVCC
VCC	CVCCC	CCCVCCC
VCCC	CVCCCC	CCCVCCCC
VCCCC		

Some languages permit even longer consonant clusters. Languages vary as to their finite inventory of segmental phonemes, as to the restrictions on specific phonemes occurring in specific positional slots in the canonical forms, and, even more important, as to the restrictions on co-occurrence in adjacent slots. There are over 2,500 languages in the world; each has from 15 to about 70 segmental phonemes; each favors certain syllable patterns; probably no two languages have the

same set of phoneme co-occurrence restrictions: This is the linguistic reality underlying the intuitive and pragmatic practices of counting letters and letter sequences in languages used by people who have by cultural accident adopted an alphabet and become literate.

"Words" are composed of one or more CVC sequences. But language is never simple. Syllables are "open" (no consonant phoneme at offglide position) or "closed" (with a consonant phoneme at offglide position). Then, too, there are the non-segmental or so-called suprasegmental phonemes (stress, pitch, tone, length, etc.) which have to be considered in the syntactic arrangements commonly called "words," idioms, or possibly phrasal nouns (the mayor of New York's hat), and utterances. E.g., the canonical form of the English monosyllabic word uttered as an exclamatory "sentence" can be represented as CVC⁴¹ (Fire! Help! Ouch! Ugh! Scram!).

The composition of words of the various languages varies as to number of syllables and as to the permitted suprasegmental patterns. Analysis shows that although words vary from mainly monosyllabic or disyllabic (as in Chinese) to mainly polysyllabic (7 or more syllables as in Hungarian or Eskimo) for a given language, still the syllable patterns fall into three major classes: utterance or word initial, medial, and final. Each syllable class in a given language is subject to specific restrictions. A five-syllable word could theoretically have the canonical form CVC-CVC-CVC-CVC-CVC: with maximal consonant phoneme clustering of 3 elements the expanded canonical form of the word would be (minus parentheses for options):

CCCVCCC-CCCVCCC-CCCVCCC-CCCVCCC-CCCVCCC.

Now let's take some language X which permits maximal consonant clustering only at word initial and word final positions and no clustering in the margins of medial syllables; the shape of the word would now be: (C)(C)(C)V(C)-(C)V(C)-(C)V(C)-(C)V(C)-(C)V(C)(C)(C). Now let us assume each word bears one and only one lexical stress. The canonical forms of the 5 syllable word in language X are: CVC-CVC-CVC-CVC-CVC, CVC-CVC-CVC-CVC-CVC, CVC-CVC-CVC-CVC-CVC, CVC-CVC-CVC-CVC-CVC, and CVC-CVC-CVC-CVC-CVC. And each of these can be expanded to the option formulations and these can then in turn be "exploded" vertically to become formulas covering the phoneme selection restrictions for each position and co-occurrence possibility.

Looking again at the 5 different patterns for the 5-syllable canonical word it can be seen that there are 6 syllable types: initial stressed (Ī), initial unstressed (I), medial stressed (Ī), medial unstressed (M), final stressed (Ī), and final unstressed (F). Let x equal any syllable type occurring more than one time in a word. An alternate canonical

symbolization would then be: $\dot{I}$ -M-x-x-F, I- $\dot{M}$ -M-x-F, I-M- $\dot{M}$ -x-F, I-M-x- $\dot{M}$ -F, I-M-x-x- $\dot{F}$.

The linguist investigates all the syllable types; the canonical forms for syllables, morphemes, and words; sentence types; and the preferences for certain patterns, as well as making finite inventories of the phonemes, affixes, and function words along with most frequent roots and the beginnings of a lexicon. Much of this information is compressed in formulaic expressions and presented as a linguistic sketch or linguistic description of a language. The basic scientific sketch and a selection of items from the lexicon can be used for a variety of applications: constructing writing systems, writing reference grammars or courses in the language, translating religious material, analysis of style, social judgments about language usage, investigating mental disorders, machine translation, cryptology, etc.

This discussion so far has merely made explicit some of the principles that underlie what was commonly known as philology in the XIXth Century and is known today as historical reconstruction. Modern day philologists (like Dell Hymes at Yale) use electronic hardware to sort the linguistic data and set up most of the comparisons. Given the hindsight we now have, practically all of the comparisons made by linguists in establishing the Indo-European language family and the relationships between the languages of the family in the past 130 years could be done (after key punching) by modern equipment in a few minutes (and printed in a few hours).

The problem is, of course, far more complicated as we come to terms with speech (as contrasted to derivatives of speech such as writing) and try to devise methods whereby machines can be programmed to "interpret" speech or perform sorting operations on running speech. The machine will have to handle (in addition to segmental phonemes, canonical forms, restriction rules, preference selections, etc.) the complexities of the suprasegmentals (stress, pitch, intensity, tone, etc., as well as psycholinguistic correlates such as "tone of voice," voice-printing, and the like). This is all extremely complicated—but recall that any and all of the patterned habits of a given language are thoroughly internalized by "any 6 year old native speaker" of a given language.

SOME CANONICAL FORMS IN SPECIFIC LANGUAGES

Most of the discussion so far has been very general and rather "ivory-towerish" ("theoretical" in the lay sense of the word), but can be clarified in some respects by examples from "real" languages.

Fijian¹ is an example of a language with a very simple set of canonical forms for the word, namely: $\# C_1 V_1 C_2 V_2 (C_3)$; $\# C_1 V_1$

¹[1], p. 236.

$C_2 V_2 C_3 V_3 (C_4)$; V ; $V_1 CV_2$; $\# CV \#$; and $\# C_1 V_1 C_2 V_2 \#$ [where $\#$ is word juncture, the C's indicate any consonant (or none), the V's any vowel, and parenthesized C is lacking when word juncture or initial consonant of a following word follows immediately].

An Indonesian language² is an example of a language with a fairly large number of morpheme canonical forms (90) and a fairly average number of segmental phonemes (6 V and 21 C phonemes). The distribution of the canonical forms is as follows:

6 canonical forms for morphemes with 1 V
24 canonical forms for morphemes with 2 V's
60 canonical forms for morphemes with 3 V's

If all C and V exponents could occur in every position there are millions of theoretical possibilities for the construction of morphemes. A structural analysis shows there are negative rules excluding certain sequence combinations and positive rules favoring certain other combinations. Inspection of a comprehensive dictionary reveals that this language actually has about 20,000 morphemes (not millions). A statistical survey of the morpheme occurrences matched against the existing canonical forms shows that 3 canonical forms (CVCVC: 6354 occurrences, CVCCVC: 3453, and CCVCVC: 1789) account for about 60 % of the morphemes, and that the inclusion of 7 additional canonical forms (CVCV: 1307, VCVC: 1240, VCCVC: 819, CVCCV: 733, CCVCCVC: 713, CVC: 636, CVCCVC: 479) raises the count to about 90 %.

Any one canonical form has a theoretical maximum number of morphemes it can generate; this number is sharply reduced after application of the rules of exclusion and of frequency preference; and only a fairly limited number of the remaining possibilities are actually "occupied" by morphemes. In this language only two of the canonical forms have high realization percentages (CVC: 76 % and CV: 48 %), but note that both patterns have rather limited maximum theoretical possibilities and that both exemplify "laws" of least anatomical effort. If there were no restrictions, the CVC pattern would generate 2,646 morphemes; with restrictions it could generate about 850; but only 636 actually occur. In percentages, about 67 % of the possibilities are eliminated by rule application and only 76 % of the remaining possibilities are realized; i.e., about 24 % of the theoretical possibilities.

The most favored canonical form in this language is the CVCVC; morphemes of that pattern account for almost one third of all the morphemes. And yet, this pattern is exploited to a very low degree. The pattern could generate over 333,000 morphemes; some 107,000

²[2], p. 248.

after rule application. Thus the favored CVCVC pattern realizes only 6% (6,354 items) of its possible morphemes.

The enormous number of non-utilized morphemes in any language shows that all languages are capable of enormous lexical expansion as the people using these languages adopt literacy and develop into complex industrialized societies. It also indicates why lexical items borrowed from other languages are re-shaped to fit into gaps of the existing patterns rather than borrowed *in toto* with alien patterns and alien phonemes. This is the linguistic basis underlying the recognition of various types of "foreign accents"; given all of the linguistic data for all languages of the world the linguist (or a software program) could predict the pronunciation of "sputnik" or "baseball" in all languages as such terms are borrowed.

To illustrate the restrictions on the selections of consonant phonemes which cluster at syllable margins let us examine the identical canonical form [CVC → (C) (C) (C) V (C) (C) (C)] used for non-inflected monosyllabic words in two related Germanic languages, English and a cognate Scandinavian language.³ There are numerous ways of analyzing the segmental consonant phonemes of a language. The analysis shown in Fig. 1 is convenient for the purposes of this paper.

	PHONEME CLASSES		
	Resonants	Fricatives	Stops and Affricates
Cognate	r l m n ŋ	f s š h j v	p t k c b d g
English	r l m n ŋ w j	θ f s š h ð v z ž	p t k č b d g ĵ

Fig. 1.—Segmental Consonant Phonemes.

Members of the resonant, fricative, and stop classes of phonemes combine in two-and-three-unit pre-vocalic and post-vocalic clusters. A mapping of the class combinations on the canonical form yields the mirror image box diagrams of Fig. 2.

Now let us look at some of the general restrictions on the consonant phonemes. In both English and the Cognate language /h/ is always prevocalic and /ŋ/ postvocalic; the Cognate's /c/ is always prevocalic and the English /ʒ/ was always postvocalic until quite recently (the name of Zsa Zsa Gabor has changed the phonemic distribution). The

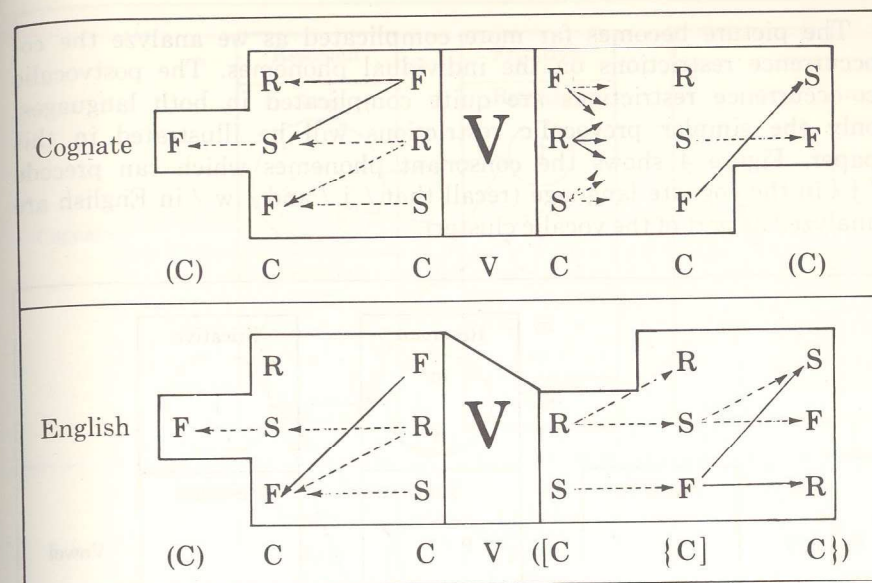


Fig. 2.—Phoneme-Class Clusters.

general restrictions on clusters are as follows. In the cognate language prevocalic /š, h, c/ never occur as members of clusters. In English prevocalic /ð, v, z, ž, c, ĵ/ never occur as members of clusters. The phoneme positional distribution is shown in Fig. 3 (positionally unique phonemes are underlined). Note that the restrictions on clustering are all prevocalic.

	PREVOCALIC		POSTVOCALIC
	Clustering	Non-Clustering	Clustering
Cognate	r l m n		r l m n <u>ŋ</u>
	f s j v	š <u>h</u>	f s š j v
	p t k b d g	<u>c</u>	p t k b d g
English	r l m n		r l m <u>ŋ</u>
	θ f s š <u>h</u>	ð v z ž	θ f s š ð v z ž
	p t k b d g č ĵ	č ĵ	p t k b d g č ĵ

Fig. 3.—Phoneme Position Distribution.

³ [3], Fig. 64, p. 132.

The picture becomes far more complicated as we analyze the co-occurrence restrictions on the individual phonemes. The postvocalic co-occurrence restrictions are quite complicated in both languages; only the simpler prevocalic restrictions will be illustrated in this paper. Figure 4 shows the consonant phonemes which can precede /j/ in the cognate language (recall that /j/ and /w/ in English are analyzed as part of the vocalic cluster).

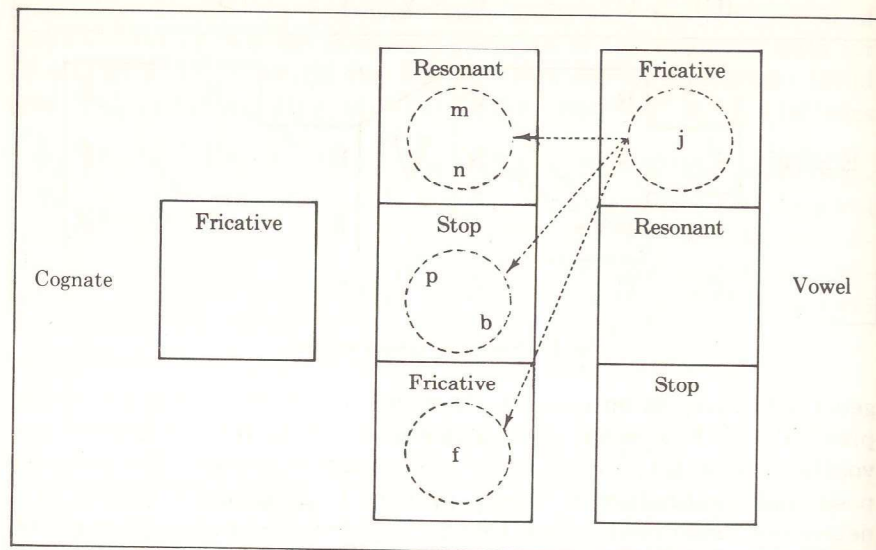


Fig. 4.

Figure 5 shows the consonant phonemes that can precede the nasal and lateral resonants (/n,m/ and /l/).

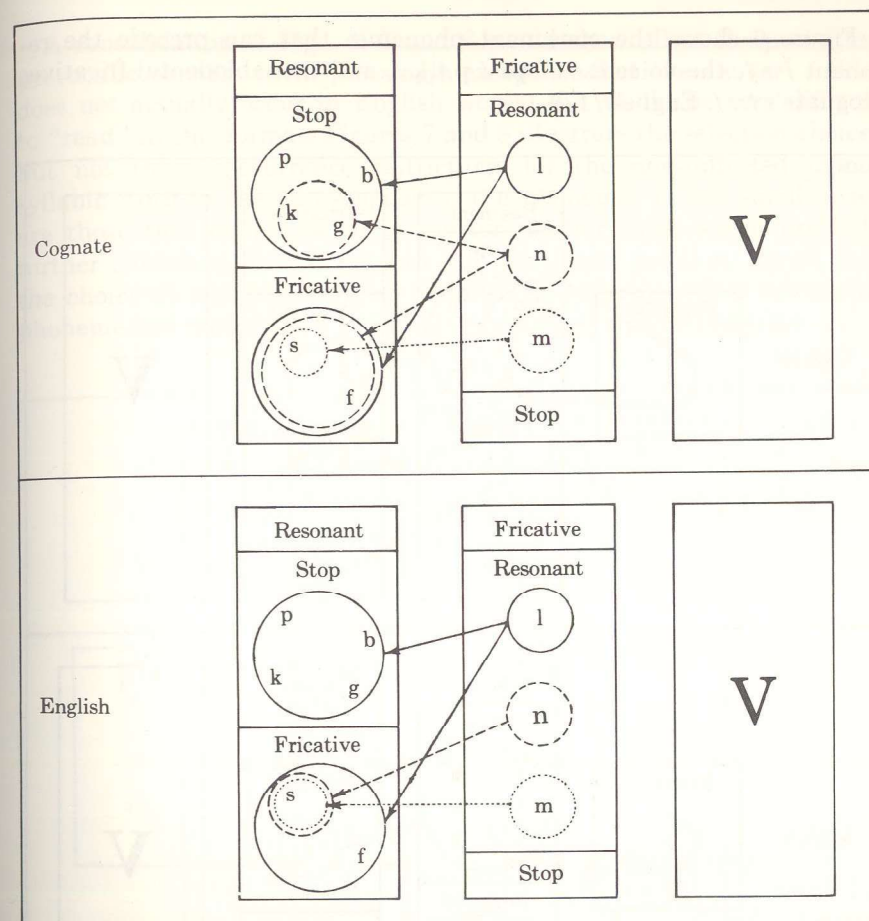


Fig. 5.

Figure 6 shows the consonant phonemes that can precede the resonant / r /, the voiceless stops / p, t, k /, and the labiodental fricatives (cognate's / v /, English / f /).

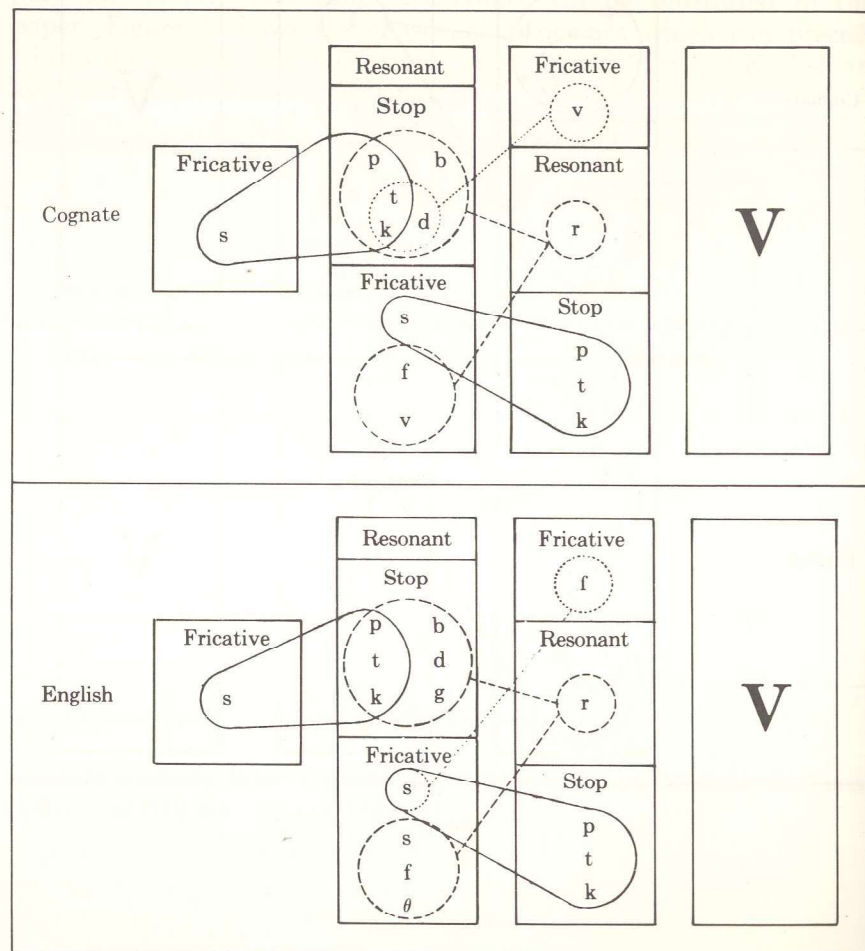


Fig. 6.

A consolidation of Figs. 4, 5 and 6 would show all of the permitted prevocalic clusters (with the single exception that the cluster / skl-/ does not actually occur in English words) but would be very difficult to "read" in this format. Figures 7 and 8 illustrate the selection choices but not the co-occurrence restrictions for the non-inflected monosyllabic word in the two languages; the phonemes in the small boxes are those that do not occur in clusters and/or those which preclude further clustering by occurring in that particular position. Recall that the choice at any position also includes the selection of no consonant phoneme and excludes the possibility of doubling a given phoneme.

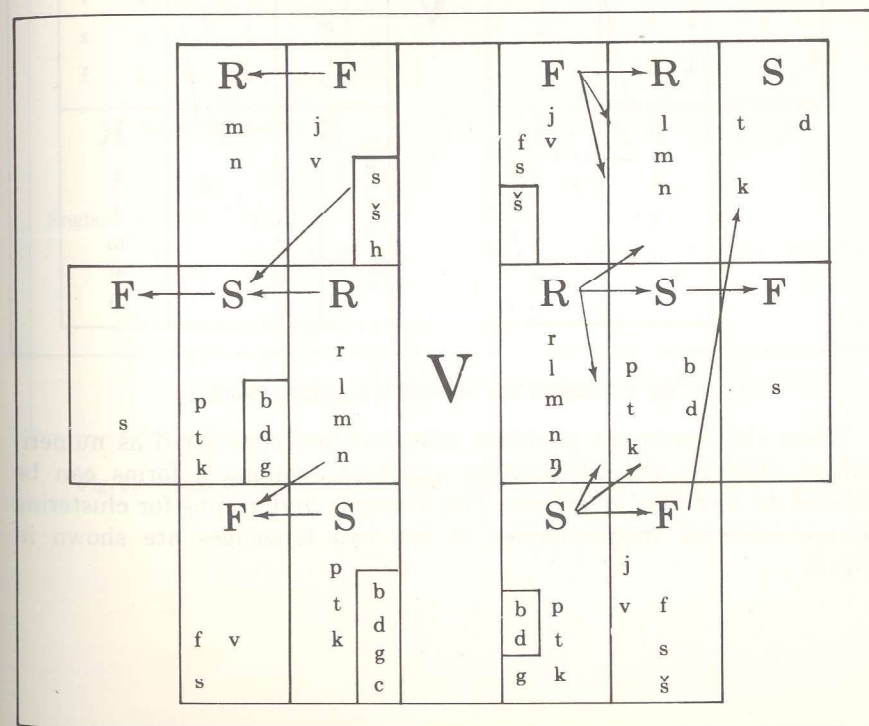


Fig. 7.—Cognate's Non-inflected Monosyllabic Word.

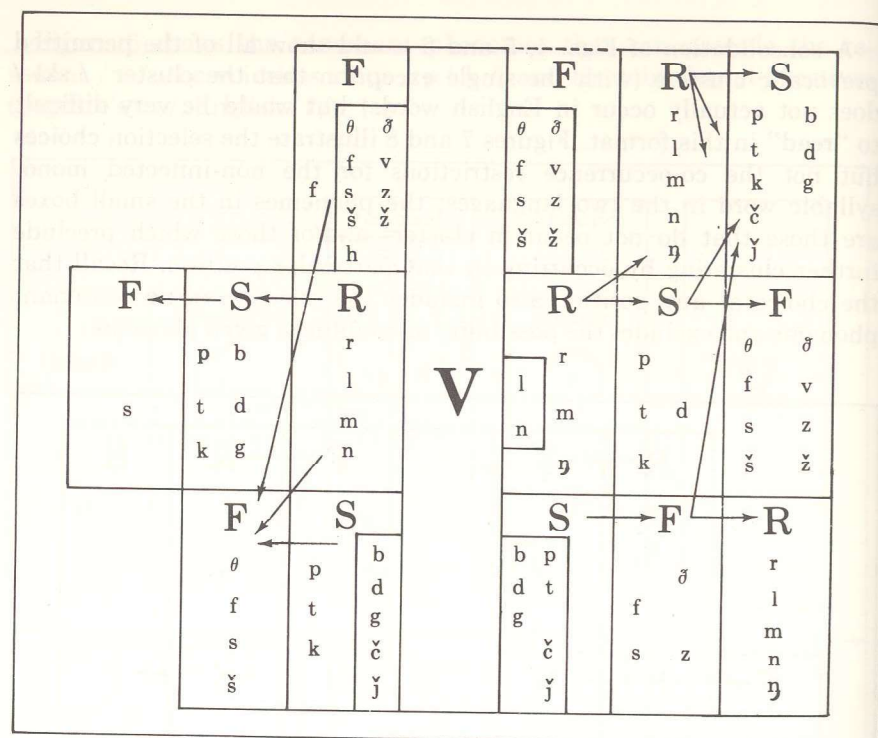


Fig. 8.—English Non-inflected Monosyllabic Word.

When the consonant phoneme selections are considered as numerical possibilities, the information about the canonical forms can be posited as formulas or graphs. The formulas and graphs for clustering in non-inflected monosyllables in the two languages are shown in Fig. 9.

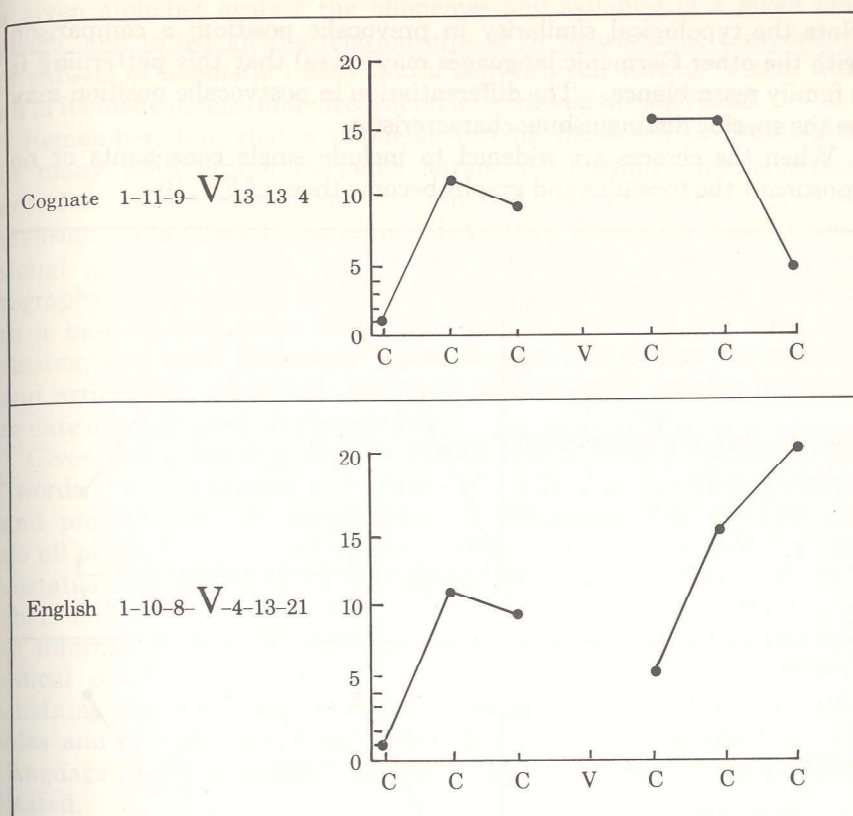


Fig. 9.—Phoneme Possibilities in Clusters Within Monosyllabic Canonical Forms.

Note the typological similarity in prevocalic position; a comparison with the other Germanic languages may reveal that this patterning is a family resemblance. The differentiation in postvocalic position may be the specific distinguishing characteristic.

When the choices are widened to include single consonants or no consonant the formulas and graphs become those of Fig. 10.

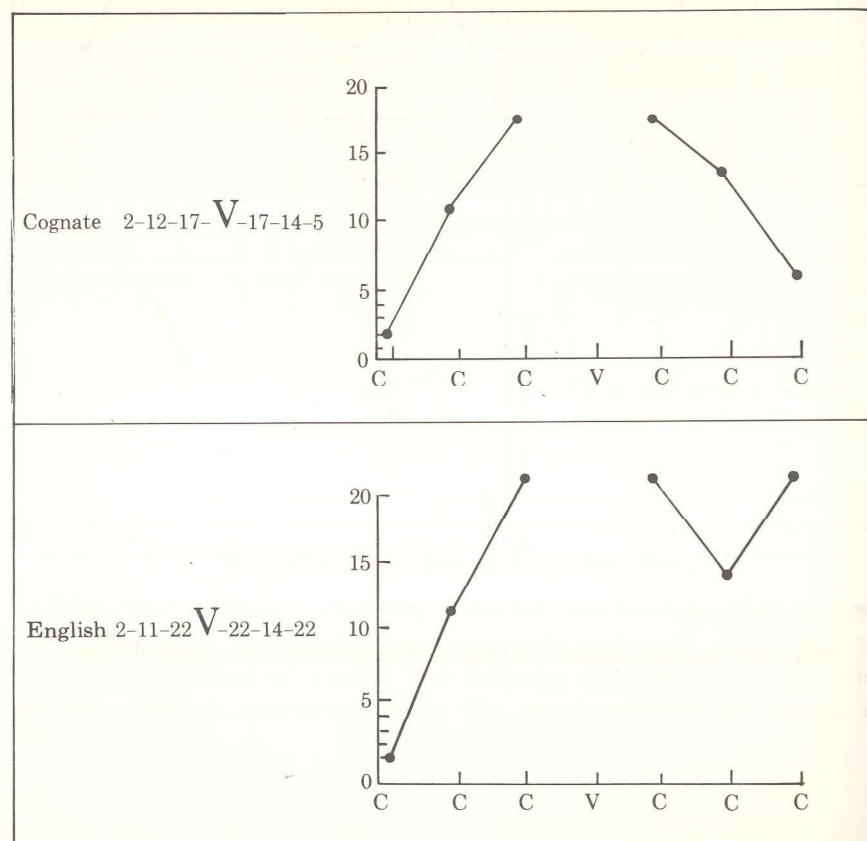


Fig. 10.—Phoneme Possibilities Within Monosyllabic Canonical Forms.

USES OF CANONICAL FORMS

How can such arcane and esoteric linguistic studies be used? Of what, if any, practical value are they or could they be? Various intermediate steps are needed before any concept or generalization of pure or theoretical science can be utilized as a technique in applied science or technology, and then, of course, still more steps before application by wider groups of specialists, technicians and analysts. To span the gap the linguist is obliged to map the written letters of

a given alphabet against the phonemes and syllables of a given language, and to list all the relationships of fit between the components of subsets of linguistic theory and a conventional writing system with all of its inadequacies and ideosyncracies garnered through time.

Remember, too, that a phoneme can be presented or represented in many ways (graph phoneme letters; oscillograms; mathematical notation as formants, formulas, numbers, or intersections; two-dimensional chartings as acoustic "distinctive features"; three-dimensional graphics; kinesic symbols for articulatory movements; palatographs; x-ray motion pictures of articulatory movements; numbers to a base of 6 or some other base for key punching and data processing; and even brainwave patterns—and this is just for acoustic and articulatory phonology inasmuch as practically nothing is known to date about perceptual phonology).

Given this "mapping" by the linguist and a set of instructions, the "words" of a dictionary or a corpus of traffic can be coded, punched and programmed for manipulation by computers. The machine can do all of the laborious (for humans) sorting tasks and provide the lexicostatistical data based on the syllable and lexeme patterns which is then used by language and communications technicians in their tasks of information processing. When the statistics are added to the canonical patterns, the material is ready for exploitation by Markov chaining and stochastic processes. The canonical forms of the syllables and of words and other lexical units can be established for any language that is of research interest and all of the restrictions precisely stated.

The techniques for the exploitation of lexicostatistical and patterned data will not be covered in this article. Linguistic analysis can have many applications; the remainder of this paper will cover techniques of linguistic analysis and proposals for application of same that will aid those language technicians who process what is known as 'voice work'.

THE NEEDS OF THE "VOICE LINGUIST"

A linguistic analysis of a language is extremely important to the "voice linguists" working as transcribers, gisters, and transcriber supervisors. What the experienced voice man believes to be his "ear" is actually many things, including aural acuity, ability to filter out noise and to disentangle two or more speech signals which at times override each other, an excellent command of the foreign intonational patterns and audible paralinguistic features of the target language, an extensive vocabulary and/or thorough familiarity with a more stereotyped smaller inventory of words and expressions, work experience, and his "feel" or "intuition" as to what he hears when the tape quality is poor and the "words" include items that are not in his active or passive vocabulary. This "feel" or "intuition" which helps him reconstruct sounds and then guess at spelling so that he can then use a dictionary or specialized glossary reflects the degree to which he has internalized the systemic relationships of the phonemes, canonical forms of syllables and morphemes and words, intonational patterns, and the statistical probabilities of same. A countering influence or cause for error is based, among other things, on the dominant internalized systemic patterning of his native speech normalized in the perceptual adjustments he is accustomed to making as he listens to speakers of other regional or social dialects (plus competing patterns from the X other languages in which he has acquired a speaking or comprehension skill). The techniques of applied contrastive linguistics can be used to isolate many of these variables and focus on the causes for indeterminacies.

But for the phoneme and syllable analysis to be of value to the "voice linguist" still another analysis must be done by the scientific linguist and presented as a chart or some other type of linguistic aid. The phonemes are the bundles of acoustic signals which mean nothing but which serve to differentiate meaning—and an alphabet is an imprecise recognition of some of these differential clues. The voice man at the skill or non-cognitive level is not dealing with letters (or words or meanings) but with the selected features, used within the linguistic system of a given language, of pulses of sound waves. Now we'll have to introduce one additional term of linguistic science jargon, that of the "allophone."

The phoneme is an early invention of mankind (indeed, for some theorists the defining characteristics of man are his oral "manipulations" of the phoneme and the applied uses of the opposable thumb)—although the term itself and the degrees of precision as to its usage in linguistic theory only goes back a few decades. But, the phoneme is an abstraction—a basic building block of language and speech, but still and all an abstraction used only by human beings with human

brains. Note how limited are the other so-called languages that are based not on an abstraction but on recognizable physical features (or derivatives of some grammatical subset of language such as numbers), e.g., the "language of flowers," the "language of gems," candy marking codes, labels for wedding anniversaries, beacons, flag systems, salutes, and even such things as apparel to mark the "in" group, or disfiguration (tattooing, scarification) or application of pigments (cosmetics) to the skin. Higher systems of communication are secondary derivatives of speech (talking drums, Canary Island whistle speech, writing systems) or tertiary derivatives (Morse alphabet, semaphore alphabet, sign language of the deaf, Braille, etc.).

The manifestation in the real world of physics of the abstraction called the phoneme is the allophone. Human sounds are just sounds; human speech sounds are allophones representing the abstract phonemes of some specific human language. A phoneme can be considered as a mental target which is aimed at by the neuromuscular speech mechanism but whose realizations are never precisely on target and never precisely the same for any repeated attempt. A convenient metaphor is that of a burst of buckshot on a paper target (Fig. 11).

Let the entire buckshot burst represent the phoneme /p/ with the target center being the mental construct and the individual pellets being the actualized [p] allophones by one individual in various /p/ attempts or the various results of many individuals in similar attempts (note that slant lines are used to mark phonemes and square brackets to mark allophones in phonemic notation). The voiced counterpart of /p/ is /b/, and the nasal voiced counterpart is /m/; all three phonemes share a number of physical correlates and each has some distinctive characteristics which set it off from the other two or at least one of the set. Now let us fire the gun three times at the target, one time for each of the three phonemes, and observe the shot displacement (Fig. 12).

See how trouble is arising. Certain acoustic phenomena can be interpreted according to patterning, "meaning" and noise interference as an allophone of one, two, three (or possibly even more) phonemes. What is he selling? A "bear" or a "pear," or is it a "mare"? Acoustically there is very little difference in the sound of the stop consonants (actually none since stop phonemes are a lack of sound with the signals in reality being carried by part of the marginal components of adjacent vocalic sounds), so maybe he said "dare," "tare" or "nair," or was it "care" or maybe "hair" (or "tear" or "hare," or even "there" as pronounced in some dialects). The brain deals and traffics at the cognitive level in "meanings" and tries to make a perceptual "Gestalt" with a meaning for any unknown series of human speech sounds. To prove this, listen to any unknown language and you will "hear" an

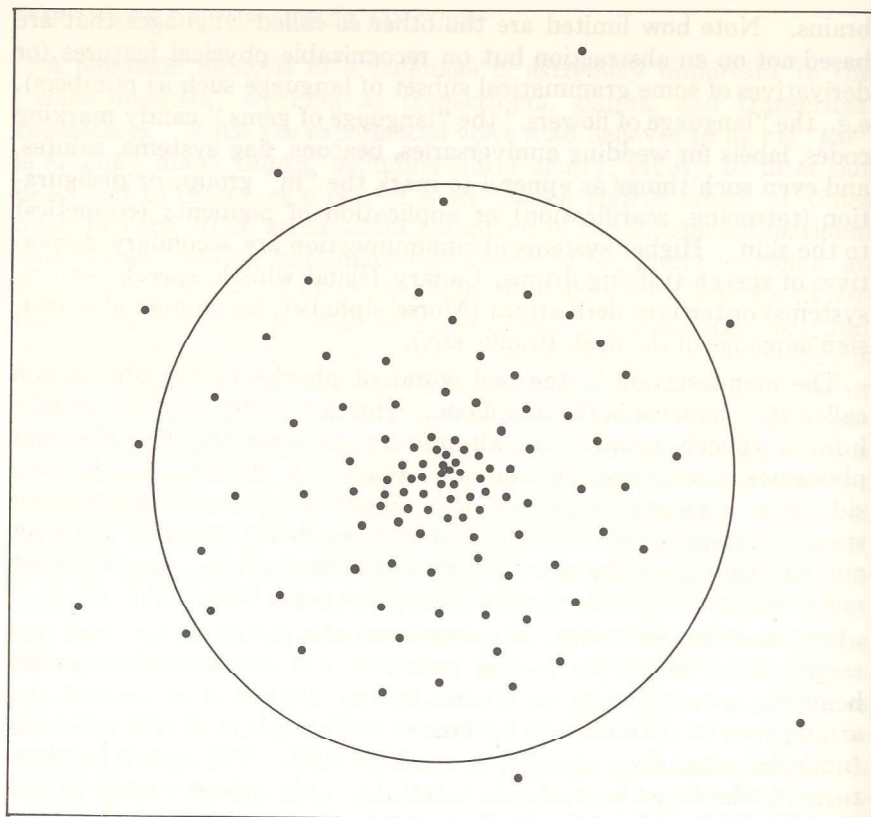


Fig. 11.—Manifestations of a Phoneme.

occasional real English word in this stream of syllabic nonsense. The mind will try its best to pattern these strange sounds into something understandable; like the many Frenchmen in World War I who were subjected to the American doughboy joke of being greeted in fluent Americanese with a "Barndoor manure!" and who wondered why those Americans laughed so hard as he politely replied with a "Bon jour, Monsieur."

To add to the complication of the phoneme concept, the sound features and buckshot pattern of a given phoneme vary according to position within a linguistic unit and according to the adjacent and even nearby sounds; e.g., the "p" in "pit" and "spit" are appreciably different allophones of the phoneme /p/ in English (but we native speakers never notice this difference until it is called to our attention). The Hindi speakers hear these two allophones of /p/ as allophones of two separate Hindi phonemes; when the Hindi speaker speaks English he will vary in his selection of phonemes to cover English words and

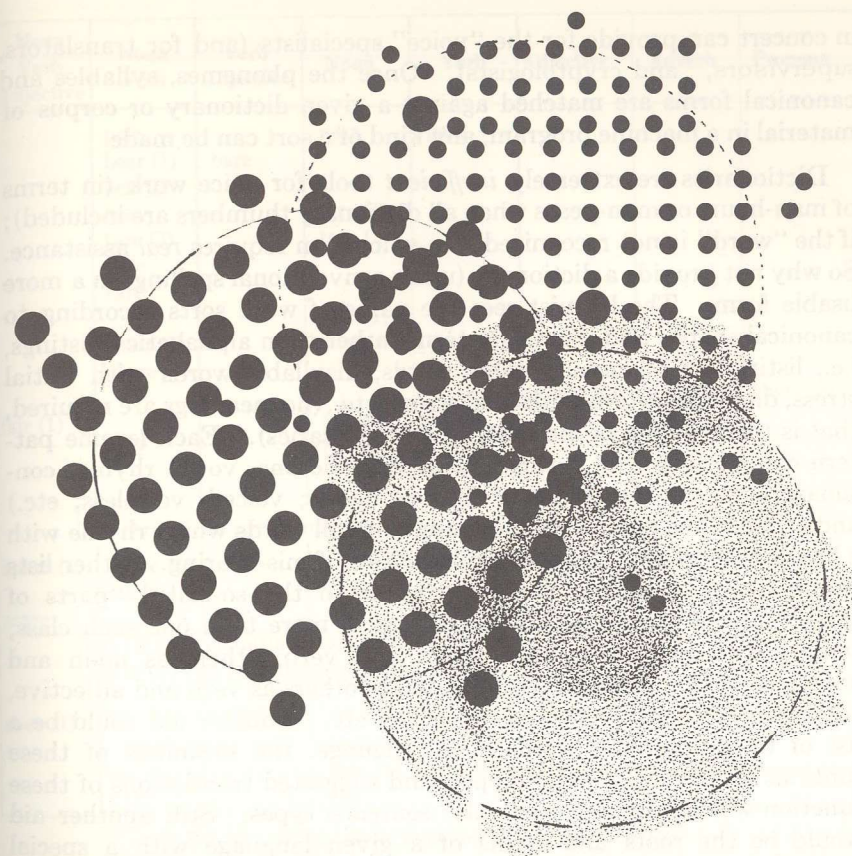


Fig. 12.—Manifestation of Three Similar Phonemes.

his attempts at "pit" and "spit" will sound to English native speakers to be "bit" and "sphit" (with variations too for the /s/, /i/ and /t/). Therefore one linguistic aid for the voice linguist would be "letter charts" representing the phonemes and distributional allophones of the target language; a variation of this aid or a separate one would be a phoneme by phoneme comparison of the sounds of English and the target language. Still other aids would be tables of statistical probabilities of occurrence of certain sounds and sequences of sounds. Other charts could cover the most frequent misinterpretations for certain sounds ("f" and "s" give difficulty in practically all languages).

The tools of the voice linguist to date are his dictionaries, glossaries, reverse sort dictionaries, rhyming dictionaries, card files, and electronic retrieval systems (information or lexical storage). In addition to the type of aids outlined in the preceding paragraph there is a whole battery of new aids and tools which linguists and programmers acting

in concert can provide for the "voice" specialists (and for translators, supervisors, and cryptologists). Once the phonemes, syllables and canonical forms are matched against a given dictionary or corpus of material in a machine program, any kind of a sort can be made.

Dictionaries are extremely *inefficient* tools for voice work (in terms of man-hours or man-years when all dictionary thumbers are included); if the "word" is not recognized, the voice man requires *real* assistance. So why not provide a dictionary (using conventional spelling) in a more usable form. The linguist sees the value of word sorts according to canonical shapes and lexeme patterns rather than alphabetical listings, i.e., listings of all monosyllabic words, disyllabic words with initial stress, disyllabic words with final stress, etc., (no meanings are required, that is the function of dictionaries and glossaries). Each lexeme pattern can have several listings: an alphabetic one, vowel rhyme, consonant class criteria (stops, fricatives, nasals; voiced, voiceless; etc.) and others, thus putting in close proximity all words which rhyme with a linguistic ordering based on probabilities of mis-hearing. Other lists would be by syntactic classes (similar to the so-called "parts of speech") or by words which participate in more than one such class; some words function as both noun and verb, others as noun and adjective, others as adjective and adverb, others as verb and adjective, still others as noun, verb and adjective, etc. Another aid could be a list of the "function" words of a language, the meanings of these function words in various sentence types, and suggested translations of these function words in those particular sentence types. Still another aid would be the roots and stems of a given language with a special section or another aid for the affixes and their permutations on the stems. An aid combining syntactic and phonological criteria for words could be devised (Fig. 13). Such aids or lists could be revamped to reflect usage statistics for various types of content or traffic.

SPEED PHONEMICS

The discussion to this point has been concerned with linguistic units as static entities rather than as active malleable units in the stream of speech. What changes do the allophonic representations of the phonemes undergo as speech speed is increased? Do some phonemes just disappear? And if so, which ones and under what circumstances? A field which has barely been scratched is the cross-disciplinary field of speed phonemics (which requires work by linguists, acoustic

Noun Verb Adjective	Noun Verb	Verb Adjective	Noun	Verb	Adjective	Adverb	Pronoun		
fair (1)	pair	bare	pear	pare	gair (2)	there (1)	there (2)		
	bear (1)							fair (3)	there (3)
	bear (2)								
	tear								
	tare (3)								
dare									
care									
	chair (1)		gair (1)						
	chare		chair (2)						
	hair		fair (2)	hare (2)					
	share		hare (1)	hare (3)					
[etc. through all initial single phonemes]									
spare (1)	spare (2)								
scare	stare		stair						
[etc. through all initial consonant clusters]									
[Then move on to next vowel nucleus]									
peer (1)			beer (1)	peer (2)					
			beer (2)						
			bier						

Fig. 13.—Word List by Function and Rhyme Scheme.

phoneticians, perceptual psycholinguist researchers, physicists, mathematicians, and hardware technicians); and little will be done until funds are made available and the need for such research realized.

Phonemes are isolated early in a linguistic investigation; the linguist then devotes his attention to the morphology, syntax and lexicon of a given language. Phonemes are elicited by comparing (minimal) pairs of lexical items or very short phrases spoken distinctly in isolation; these phonemes are normally presented as graphic symbols on paper and little further investigation is carried out unless the language concerned is characterized by "interesting" or complicated morpho-phonemic patterns. An example in English of such interesting patterns is shown in (the pronunciation and spelling of) the following word pairs:

knife: knives	child: children	ride: rode
foot: feet	wide: width	sheep: sheep

Speech and language permit communication despite considerable suppression of acoustic signaling data through the devices of syntactic,

semantic and phonetic redundancy; thus we can understand young children, toothless oldsters, gum chewers, pipe biters, and even shouters in a boiler factory. This is true for all languages. In literate societies, suppression of phonetic data by the speaker (as contrasted to environmental noise) is judged to be "sloppy" speech; the prescriptive standard or imagined "norm" is the written form of the words and words are supposed to be enunciated as they are written (whatever that means). The normative prescriptive teacher of English tries to make the students make a differentiation in their pronunciation of "sentences" like the following ("phonemic" notation that of a dictionary and not truly phonemic):

Picture right here!	'pikchə(r)
Pitcher right here!	'pichə(r)
Pitch her right here!	pich (h)ər

Most people *can* make a differentiation in slow studied speech; but in regular or rapid speech there is no difference in the above three examples [pichə(r)]. The semantic differentiation is supplied by the situational context, not by the phonemes and not by the syntax; the "words" are an after-the-fact analysis supplied by the hearer.

The common languages taught in America have been French and Spanish; and Spanish is "judged" to be easier to learn. But is it? The judgement is really concerned with phonology and the spelling conventions. "Spanish is phonetic" (or "Russian is phonetic"); it sounds the way it is spelled! In reality, Spanish is a syllable-timed language and such languages are characterized by far less phonetic data suppression than are stress-timed languages such as English; thus Spanish is "easy" for English speakers. Maybe the Martians "speak" with their eyes; thus their language is not phonetic; but we gaeans (or terrans), unless mute, all speak phonetically.

French, on the other hand, is a breathgroup final-stress language characterized by considerable phonetic suppression at the beginning of breath groups. This factor, plus a fairly "irregular" spelling convention, underlies the judgement that French is "hard." Teachers of French practically all start out with phonetic rather than spelled representations of the words and phrases and later use these symbols to show the regular changes that occur in actual speech (permutations and replacements). Some examples are:

je ne sais pas	[ʒənəsɛpa]	→	[ʒənsɛpa]
je n'irai pas	[ʒənirɛpa]	→	[ʃnirɛpa] or [ʒirɛpa]
je m'en vais	[ʒəmāvɛ]	→	[ʃmāvɛ]
je viens	[ʒəvjɛ]	→	[ʃvjɛ] → [ʃfjɛ]

The same factors operate in English. Following are some examples of "sloppy" (i.e., rapid "colloquial" speech) English:

Formal	Colloquial	Phonetic renditions
Let us go eat.	Let's go eat.	lɛzgəwiyt → lɛsgwiyt
Did you see him?	Did'ya see 'im?	dɪʃəsiyim → ʃusiym
Come here!	C'm'ere!	kmiyr
Stop complaining!	Quit your bitching!	kwičə-bičn
John is here!		ʃanzɪyɪr ~ ʃansɪyɪr ~ ʃanʃɪyɪr

The last example can be interpreted as: John's here; John's year; John's ear; John's cheer; John's shear; The john is here; John's sear; John's seer; etc.

Speed phonemics research would have application in voice transcription, voice transcription aids, voice transcription training, speech ciphony and other forms of artificial speech or speech-by-rule, and in the electronic processing of recorded speech matter (voice-printing, language identification, machine transcription). When the stream of speech rather than the syllable patterns within words is examined, changes occur in the permitted clustering of phonemes (changes of composition and changes of frequency). For example, the initial cluster tm- is rare in English words (tmesis), as are tl- (Tlingit), gw- (Gwen), kn- (knockwurst) and ts- (tsetse), but are quite common across word boundaries in running speech (t'morrow; out'late; "g'wan now"; C'n I go?; 't's all over now). Indeed there are even some new s- initiated 3-phoneme clusters (stl- as in "it is too late" rendered as s't'late; for some people a glottal stop can precede the s-, this forms a 4 part initial cluster). Similar changes in elision, compression, and quality occur with the vowel phonemes; compare the similar feature in English and Russian whereby unaccented vowel phonemes are reduced to schwa.

Should I stress what is now the obvious—that the "system" of conventions covering hyphenation of spelled English words has little correlation to syllable boundaries in speech. The hyphenation convention (with its origins in the field of typesetting and subsequent development by editors) is roughly based on morpheme boundaries within words along with a series of ad hoc "rules" applied non-consistently. Compare a few words hyphenated according to the "rules" as shown in a dictionary and according to real speech:

know-ing	[no-(w)ɪŋ];	knowl-edge	[na-liʃ];	ac-knowl-edge	[æ k-na-liʃ or ə-kna-liʃ];
ac-knowl-edg(e)-ment	[æ(k)-(k)na-li-ʃmint];	lan-guid-ly	[lɛŋ-gwi-dli or lɛ-wi-dli];	lank-i-ly	[læ ŋ-ki-li or læ ŋ-kli]

The above comparison leads to an explanation of the fact that in reading it is sometimes necessary to back up and re-scan a hyphenated word—the incorrect phonemic signal was projected by the spelling of the portion of the word preceding the hyphen, as in the following examples:

Have you seen my lan-

{ guage book?
{ tern?
{ ky brother?

Then he picked up the mus-

{ cadine grape.
{ cly meat.
{ ing mouse.

Let us now examine a typical sentence fragment full of short words when all possible word and syllable divisions are marked (despite the ad hoc rule about not hyphenating certain short words):

“It is easy to see why the analyst in using language . . .”

it-is-ea-sy-to-see-why-the-a-na-lyst-in-us-ing-lan-guage . . .

The same speech utterance at 4 progressively faster speeds can be represented by a combination of “eye dialect” and phonemes as follows:

its-sea-sy-to-see-wai-thə-æ-nə-lyst-in-yu-zing-læn-gwidge . . .

i-tsiy-zi-tə-siy-wai-ðə-æn-nə-lis-tin-yu-zin-læ-wiŋ . . .

i-tsiy-zi-tsiy-wai-ðæn-nə-lis-tin-nyu-zən-læ-wiŋ . . .

tsiy-zi-tsiy-wai-ðæn-li-stin-nyu-zən-læ-wiŋ . . .

There are, of course, many possible variations reflecting regional, social and ideosyncratic dialects; thus the phoneme patterns of distribution in a given language express this variation in generalized allophonic distributional sets per investigated dialect. When the factor of utterance speed is included, it is obvious that “speed phonemics” will have to take the phonological syllable as the basic unit of investigation. And any type of application of electronics or “machines” to the perception, transcription or translation of oral communication data will also have to be based on the acoustic signals—and not on written words.

FUTURE LINGUISTIC RESEARCH

An analysis as presented in this paper based on language rather than on a secondary derivative cultural phenomenon (writing) has implications for the future. Enormous sums of money have been spent in the past 15 years on mechanical translation. In the early days most of the money disappeared in attempts to mechanize lexical, morphological and phrasal concordance equivalents in some two given languages. Much of this work was not directed by competent linguists (mainly because there were very few such linguists who could also cope with the new realm of cybernetics and computers) in the initial years. The

current theoretical work and trial applications in computational linguistics and theoretical machine applications for language work deal mainly with the syntactic deep structure of languages. Theoreticians are now investigating theories of the semantic component of language. Most of this involves the devising of models and theoretical constructs which simulate, parallel, or can be said to stand in some sense for some one portion of the “black box” processes which take place in the human brain as the individual reacts and performs psycholinguistically in a sociolinguistic situation in the physical environment in which he lives.

Linguistics has been at the “lever and pulley” stage for the past century, is currently making rapid strides (thanks partly to computer technology and its challenges), and should be able to make enormous advances when and if electronic devices are designed to handle language tasks (rather than pushing and squeezing language into devices designed to perform counting tasks based on binary decision procedures).

The input and output of the devices of the present and of the future will be in the components of surface structure (syllables and phonemes “interpreted” as morphemes; derivative letters and written words reflecting or representing certain aspects of the phonemes and morphemes). In the fairly near future the electronic input from optical scanning of texts in language X will be manipulated in some series of programming procedures and be “printed” in language Y (i.e., one step more advanced than the current primitive machine translation devices which by that time should be a bit more sophisticated). The next step will be to use the primary units of speech (syllable and phoneme) and attack the problems of transcription of speech. Step three will marry the latter technology to the then existing translation technology. The first solutions will probably convert recordings of sound waves of language X into “written” forms of language Y which can be “read” or deciphered by human beings as well as fed into various types of information storage units for information retrieval as required. Later and still more sophisticated solutions will provide devices for sequential and simultaneous interpretation in two or more languages.

REFERENCES

- [1] Charles F. Hockett, “Problems of Morphemic Analysis”; reprinted in Martin Joos, ed., *Readings in Linguistics*, American Council of Learned Societies, Washington, 1957.
- [2] E. M. Uhlenbeck, “The Structure of the Morpheme”; reprinted in Hamp, Householder, and Austerlitz, eds., *Readings in Linguistics II*, University of Chicago Press, Chicago, 1966.
- [3] Bertil Malmberg, *Structural Linguistics and Human Communication*, Springer-Verlag OHG, Berlin-Göttingen-Heidelberg, 1963.

Contributors

CLIFTON H. BUTCHER—Language studies at Cincinnati, Harvard, Middlebury, Tennessee, Paris and Georgetown; Certificate *magna cum laude*, Chinese Language and Area Studies, ASTP, Harvard, 1943-44; A.B. *cum laude*, General Studies (Chinese), Harvard, 1947; M.A., French, Middlebury, 1948; course work completed for Ph.D. in Theoretical Linguistics. Translator and Interrogator, USAREUR Intelligence Center, 1951-55; Chief, Slavic Languages Branch, and Lecturer, USAREUR Intelligence School, Oberammergau, 1955-56; Training, Transcriber, Analyst, Cryptologic Linguist, NSA, 1956-. NSA Fellow, 1964-65. Author of *Polish Productive Roots Flash Cards*, 1959; *Illustrative Index of Productive Polish Affixes*, 1965; "Linguistic Canonical Forms," *NSA Technical Journal*, Vol. XIII, No. 1, Winter 1968; "Soviet Linguistics—The State of the Art," *Ibid.*, Vol. XV, No. 1, Winter 1970.

JACOB GURIN—B.S., Education, College of the City of New York, 1940; M.A., The George Washington University, 1960. Japanese translator and interpreter (personal interpreter to CG, 8th Army), USA 1942-46; NSA and predecessor agencies, 1946-; Linguistic Research, Training, Speech Research, various administrative positions. Member of CLA and of the Advisory Board of *Cryptologic Spectrum*, Chairman of the Board of Directors of The Phoenix Society. Author of "The Translation of a Classic: A Matter of Ethics," *NSA Technical Journal*, Vol. XI, No. 2, Spring 1966; "Employment of Military Linguists," *Ibid.*, Vol. XIII, No. 4, Fall 1968; Speaker Identification by Machine: Where Do We Stand?, *Ibid.*, Vol. XIV, No. 4, Fall 1969; "A Measure for Transcription Efficiency," *Ibid.*, Vol. XV, No. 1, Winter 1970.

CLYDE D. MCDONALD—B.S., Mathematics, Tulane University, 1952; Graduate, Army Language School (Russian), 1961; M.A. in Public Administration (Technology of Management), American University; ASW, USN, 1952-54; CIC, USN, 1954-55; NSA 1955-58, 1964-67; Army Language School, 1960-61; Naval Security Group, duty in U.S. and Japan, 1958-60, 1961-64, 1968-.

HARRY G. ROSENBLUH—A.B., Romance Languages, University of Pennsylvania, 1951; further language studies at NSA. NSA, 1951-: Research Analytic Specialist, Translator-Checker, Cryptologic Linguist. Editor of *Quarterly Review for Linguists*; Secretary, Production Organization Language and Linguistics Group 1966-; Secretary, Crypto-Linguistic Association, 1969-; Spiritual Leader, Jewish Congregation of Laurel, 1966-. Member of Crypto-Mathematics Institute, Crypto-Linguistic Association, Pi Delta Phi, ALMA'I (Association of Teachers of the Hebrew Language). Author of *A & P: An Inventory of Linguistic Aids and Projects*, 1967; *Nicknames and Diminutives in Spanish*, 1967; *Nationality Guide for Latin American Spanish Material*, 1967; "The Spanish Languages," *NSA Technical Journal*, Vol. XIII, No. 1, Winter 1968; "Mokusatsu: One Word, Two Lessons," *Ibid.*, Vol. XIII, No. 4, Fall 1968.

ARTHUR J. SALEMME—B.S., Russian, Georgetown University, 1954; M.S., Languages, Georgetown, 1958. Signal Corps, USA, 1943-46; Research Analyst and Linguist, NSA, 1947-; U.S. Army Institute of Advanced Soviet Studies (Detachment "R"), 1961-63; Chairman, first NSA Russian-language certification panel, 1967; current NSA assignment: Cryptologic Linguist. Author of *Distribution of Grammatical Particles in a Russian Chemical Text as Applicable to Machine Translation* (Master's thesis, in Russian), 1958; "Some Problems in the Field of Machine Transla-

tion," *NSA Technical Journal*, Vol. V, Nos. 3 and 4, 1960 (Award, CMI Essay Contest); "Are the Russians Gaining Wordwise?," *Ibid.*, Vol. VI, No. 2, 1961; *Russian Spelling vs. Russian Pronunciation* (Detachment "R" thesis), 1963; "A Few Tricks of the Trade for the Translator of Russian," *NSA Technical Journal*, Vol. XI, No. 4, 1966 (Award, CIA Essay Contest); "Handle With Care: Russian Transliteration," *Ibid.*, Vol. XII, No. 4, 1967 (Award, CIA Essay Contest); "A Tour of Some Russian Translation Traps," *Ibid.*, Vol. XIV, No. 4, 1969.

JOHN H. TILTMAN—Charterhouse, 1914; British Army, 1914-25, 1930-46, retired 1946 as Brigadier. Government Communications Headquarters, 1920-. Deputy Director and Chief Cryptographer, 1941-46; Commander, Order of St. Michael and St. George; Commander, Order of the British Empire; Military Cross.

